LES FACULTÉS
DE L'UNIVERSITÉ
CATHOLIQUE DE LILLE

Data Observation

EXPLORATORY DATA ANALYSIS

Baptiste Mokas

baptiste.mokas@gmail.com

weeki.io/dynamical-système

linktr.ee/baptistemokas

+33 7 69 08 54 19

KNOWLEDGE TREE

Part 1: Descriptive Statistics

1.1 Introduction to Descriptive Statistics

- Purpose and significance of descriptive statistics
- Overview of the exploratory data analysis (EDA) process
- Key statistical vocabulary

1.2 Understanding and Representing Data

- Data types: numeric, categorical
- Measurement scales: nominal, ordinal, interval, ratio
- Data representation techniques

1.3 Data Preprocessing and Cleaning

- Handling missing data
- Detection and treatment of outliers
- Data scaling and normalization methods

1.4 Data Visualization

- Importance of data visualization
- Common visualization types: histograms, box plots, scatter plots
- Visualization tools: Matplotlib, Seaborn.

1.5 Analyzing Relationships in Data

- Understanding correlation coefficients: Pearson, Spearman
- Exploring covariance
- Visualization techniques: scatter plots, correlation heatmaps

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.1 Introduction to PCA

- Mathematical foundations of PCA
- Interpretation of principal components

2.2 Data Representation and Summarization using PCA

- Frequency distributions: frequency tables, histograms
- Data summaries: percentiles, quartiles, box plots
- Data transformations: logarithmic and square root

2.3 Exploring Bivariate and Multivariate Relationships

- Joint, marginal, and conditional distributions
- Statistical independence and linkage
- Analysis with more than two variables

2.4 Advanced Visualization

- Introduction to advanced visualization tools: ggplot2, etc.
- Creating complex visualizations

2.5 Case Study in PCA

- Proposing and defining a PCA project
- Analysis, interpretation, and conclusions
- Presentation of findings and feedback

Part 3: Integration of Global Concepts and Advanced Topics

3.1 Data Distribution and Shape

- Probability distributions: normal, binomial, Poisson
- Measures of skewness and kurtosis

3.2 Moments and Central Moments

- Moments for characterizing data: mean, variance, skewness, kurtosis

3.3 Multivariate Descriptive Statistics

- Multivariate statistics: multivariate mean, covariance matrix

3.4 Descriptive Statistics in Big Data

- Challenges and techniques for analyzing large datasets
- Distributed computing and descriptive statistics

3.5 Practical Applications of Descriptive Statistics

- Applications in business, economics, healthcare, social sciences
- Practical problems, solutions, and case studies

Part 4: Step-by-step Construction of PCA

4.1 Understanding and Foundations of PCA

- Background, historical context, and significance of PCA.
- The need for dimensionality reduction in data analysis.
- Standardization of Data: Ensuring uniformity in feature scales.

4.2 Construction and Decomposition

- Building the Covariance Matrix: Understanding mutual variance.
- Eigen Decomposition: Grasping eigenvalues and eigenvectors to derive principal components.

4.3 Orthogonal Transformations and Interpretations

- Introduction to OCA (Orthogonal Component Analysis) and its link with PCA.
- Conversion of correlated variables into an orthogonal set.
- Projection onto new coordinates using top eigenvectors and interpreting their significance.

4.4 PCA Visualization and Applications

- Methods for visualizing PCA results: scree plots, biplots, etc.
- Real-world applications, case studies, and the benefits of PCA.

4.5 Advanced Topics and Conclusion

- Delving into advanced PCA variants: Robust PCA, Kernel PCA.
- Analyse en Composante Principale: A look into the French context and terminologies.
- Recap of PCA's pivotal role in data analysis and recommendations for further exploration.

Purpose and significance of descriptive statistics

Part 1: Descriptive Statistics

1.1 Introduction to Descriptive Statistics

Overview of the exploratory data analysis (EDA) process

Part 1: Descriptive Statistics

1.1 Introduction to Descriptive Statistics

Key statistical vocabulary

Part 1: Descriptive Statistics

1.1 Introduction to Descriptive Statistics

Data types: numeric, categorical

Part 1: Descriptive Statistics

1.2 Understanding and Representing Data

Measurement scales: nominal, ordinal, interval, ratio

Part 1: Descriptive Statistics

1.2 Understanding and Representing Data

Data representation techniques

Part 1: Descriptive Statistics

1.2 Understanding and Representing Data

Handling missing data

Part 1: Descriptive Statistics

1.3 Data Preprocessing and Cleaning

Detection and treatment of outliers

Part 1: Descriptive Statistics

1.3 Data Preprocessing and Cleaning

Data scaling and normalization methods

Part 1: Descriptive Statistics

1.3 Data Preprocessing and Cleaning

Importance of data visualization

Part 1: Descriptive Statistics

1.4 Data Visualization

Common visualization types: histograms, box plots, scatter plots

Part 1: Descriptive Statistics

1.4 Data Visualization

Visualization tools: Matplotlib, Seaborn.

Part 1: Descriptive Statistics

1.4 Data Visualization

Understanding correlation coefficients: Pearson, Spearman

Part 1: Descriptive Statistics

1.5 Analyzing Relationships in Data

Exploring covariance

Part 1: Descriptive Statistics

1.5 Analyzing Relationships in Data

Visualization techniques: scatter plots, correlation heatmaps

Part 1: Descriptive Statistics

1.5 Analyzing Relationships in Data

Mathematical foundations of PCA

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.1 Introduction to PCA

Interpretation of principal components

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.1 Introduction to PCA

Frequency distributions: frequency tables, histograms

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.2 Data Representation and Summarization using PCA

Data summaries: percentiles, quartiles, box plots

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.2 Data Representation and Summarization using PCA

Data transformations: logarithmic and square root

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.2 Data Representation and Summarization using PCA

Joint, marginal, and conditional distributions

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.3 Exploring Bivariate and Multivariate Relationships

Statistical independence and linkage

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.3 Exploring Bivariate and Multivariate Relationships

Analysis with more than two variables

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.3 Exploring Bivariate and Multivariate Relationships

Introduction to advanced visualization tools: ggplot2, etc.

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.4 Advanced Visualization

Creating complex visualizations

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.4 Advanced Visualization

Proposing and defining a PCA project

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.5 Case Study in PCA

Analysis, interpretation, and conclusions

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.5 Case Study in PCA

Presentation of findings and feedback

Part 2: Principal Component Analysis (PCA) and Advanced Visualization

2.5 Case Study in PCA

Probability distributions: normal, binomial, Poisson

Part 3: Integration of Global Concepts and Advanced Topics

3.1 Data Distribution and Shape

Measures of skewness and kurtosis

Part 3: Integration of Global Concepts and Advanced Topics

3.1 Data Distribution and Shape

Moments for characterizing data: mean, variance, skewness, kurtosis

Part 3: Integration of Global Concepts and Advanced Topics

3.2 Moments and Central Moments

Multivariate statistics: multivariate mean, covariance matrix

Part 3: Integration of Global Concepts and Advanced Topics

3.3 Multivariate Descriptive Statistics

Challenges and techniques for analyzing large datasets

Part 3: Integration of Global Concepts and Advanced Topics

3.4 Descriptive Statistics in Big Data

Distributed computing and descriptive statistics

Part 3: Integration of Global Concepts and Advanced Topics

3.4 Descriptive Statistics in Big Data

Applications in business, economics, healthcare, social sciences

Part 3: Integration of Global Concepts and Advanced Topics

3.5 Practical Applications of Descriptive Statistics

Practical problems, solutions, and case studies

Part 3: Integration of Global Concepts and Advanced Topics

3.5 Practical Applications of Descriptive Statistics

Background, historical context, and significance of PCA.

Part 4: Step-by-step Construction of PCA

4.1 Understanding and Foundations of PCA

The need for dimensionality reduction in data analysis.

Part 4: Step-by-step Construction of PCA

4.1 Understanding and Foundations of PCA

Standardization of Data: Ensuring uniformity in feature scales.

Part 4: Step-by-step Construction of PCA

4.1 Understanding and Foundations of PCA

Building the Covariance Matrix: Understanding mutual variance.

Part 4: Step-by-step Construction of PCA

4.2 Construction and Decomposition

Eigen Decomposition

Part 4: Step-by-step Construction of PCA

4.2 Construction and Decomposition

Introduction to OCA and its link with PCA.

Part 4: Step-by-step Construction of PCA

4.3 Orthogonal Transformations and Interpretations

Conversion of correlated variables into an orthogonal set.

Part 4: Step-by-step Construction of PCA

4.3 Orthogonal Transformations and Interpretations

Projection onto new coordinates using top eigenvectors and interpreting their significance.

Part 4: Step-by-step Construction of PCA

4.3 Orthogonal Transformations and Interpretations

Methods for visualizing PCA results: scree plots, biplots, etc.

Part 4: Step-by-step Construction of PCA

4.4 PCA Visualization and Applications

Real-world applications, case studies, and the benefits of PCA.

Part 4: Step-by-step Construction of PCA

4.4 PCA Visualization and Applications

Delving into advanced PCA variants: Robust PCA, Kernel PCA.

Part 4: Step-by-step Construction of PCA

4.5 Advanced Topics and Conclusion

Analyse en Composante Principale

Part 4: Step-by-step Construction of PCA

4.5 Advanced Topics and Conclusion

Recap of PCA's pivotal role in data analysis and recommendations for further exploration.

Part 4: Step-by-step Construction of PCA

4.5 Advanced Topics and Conclusion

KEYWORDS (NEW)

Variable discrète vs continue
Variable statistique continue
Variable statistique discrète
Percentiles
Moyenne empirique
Estimation, intervalles de confiance
Pourcentage cumulé
Moyenne
Multicolinéarité
Autocorrélation
Quartiles
Moyenne arithmétique
Analyse bivariée et corrélation
Skewness (asymétrie)
Tableau
Tendance centrale
Coefficient de corrélation
Co-linéarité
Diagramme en boîte (box plot)
Indépendance statistique
Statistique descriptive
Variance
Écart absolu moyen
Coefficient de variation
Écart-type
Distribution de fréquence
Indépendance statistique conjointe
Variance empirique
Asymétrie
Écart interquartile
Corrélation de Spearman
Tableau statistique relatif à un caractère qualitatif et sa représentation graphique
Matrice de variance-covariance
Etendu interquartile
Corrélation
Médiane
Mode
Tableaux statistiques et représentations graphiques
Diagramme
Dispersion
Erreur standard
Matrice de corrélation
Covariance
Minimum
Matrice de covariance
Etendu (range)

KEYWORDS

- Exploratory Data Analysis (EDA)
- Data Exploration
- Data Understanding
- Data Analysis
- Data Scientist
- Data Analyst
- Researcher
- Data Patterns
- Data Relationships
- Data Insights
- Data Types
- Numeric Data
- Categorical Data
- Data Scales
- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale
- Data Representation
- Data Formats
- Data Cleaning
- Data Preprocessing
- Missing Data
- Outliers
- Data Scaling
- Data Normalization
- Data Visualization
- Data Plots
- Histograms
- Box Plots
- Scatter Plots
- Matplotlib
- Seaborn
- Visual Narratives
- Descriptive Statistics
- Measures of Central Tendency
- Mean
- Median
- Mode
- Measures of Dispersion
- Range
- Variance
- Standard Deviation
- Interquartile Range (IQR)
- Coefficient of Variation
- Correlation Coefficients
- Pearson Correlation
- Spearman Correlation
- Correlation Heatmaps
- Data Distributions
- Normal Distribution
- Probability Density Function (PDF)
- Z-Scores
- Normality Tests
- Skewness
- Kurtosis
- Asymmetry
- Tailedness
- Categorical Data Analysis
- Frequency Tables
- Cross-Tabulations
- Bar Charts
- Pie Charts
- Chi-Square Test for Independence
- Advanced EDA Techniques
- Data Transformation
- Log Transformation
- Box-Cox Transformation
- Skewed Data
- Multivariate Exploratory Data Analysis
- Principal Component Analysis (PCA)
- Multidimensional Scaling (MDS)
- Cluster Analysis
- Hidden Patterns
- Time Series Analysis
- Time Series Decomposition
- Seasonal Decomposition of Time Series (STL)
- Autocorrelation Functions

Use Case with python

Use Case: Customer Behavior Analysis with Exploratory Data Analysis (EDA)

In the context of the course on Exploratory Data Analysis (EDA), covering topics on data understanding, cleaning, visualization, and advanced EDA techniques, let's explore a use case related to analyzing customer behavior using EDA. This use case involves applying EDA techniques to understand and gain insights from customer data.

Description:

In this use case, we will focus on analyzing customer behavior data to gain valuable insights and improve business decision-making. We will use EDA techniques to explore customer data, detect patterns, and visualize relationships between customer variables.

Key Components:

Introduction to Data Observation and EDA: Understanding the basics of data types, data scales, data representation, data cleaning, and data visualization.

Descriptive Statistics: Utilizing measures of central tendency, measures of dispersion, and exploring relationships between customer variables.

Data Distributions: Analyzing data distributions, including normal distribution, skewness, and kurtosis, and performing categorical data analysis.

Advanced EDA Techniques: Leveraging advanced EDA techniques such as data transformation, multivariate exploratory data analysis, and time series analysis.

Python Code Example (Customer Behavior Analysis):

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import skew, kurtosis, chi2_contingency

# Load customer behavior data
data = pd.read_csv('customer_data.csv')

# Summary statistics
summary_stats = data.describe()

# Distribution of customer age
plt.figure(figsize=(10, 6))
sns.histplot(data['Age'], bins=20, kde=True)
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Distribution of Customer Age')
plt.grid(True)

# Correlation matrix
correlation_matrix = data.corr()

# Heatmap of correlations
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()

# Chi-square test for independence (categorical data)
contingency_table = pd.crosstab(data['Category'], data['Purchase'])
chi2, p, dof, _ = chi2_contingency(contingency_table)
print(f'Chi-square Statistic: {chi2:.2f}')
print(f'p-value: {p:.4f}')

# Log transformation of a skewed variable
data['Log_Sales'] = np.log(data['Sales'])

# Principal Component Analysis (PCA)
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
data_pca = pca.fit_transform(data[['Age', 'Income', 'Purchase']])

```

In this code, we load customer behavior data, calculate summary statistics, visualize the distribution of customer age, create a correlation heatmap to explore relationships between variables, perform a chi-square test for independence on categorical data, and apply a log transformation to a skewed variable. Additionally, we demonstrate principal component analysis (PCA) for multivariate analysis.

This use case demonstrates how EDA techniques can be applied to customer behavior data to uncover insights, improve customer targeting, and enhance business strategies.

REFERENCES V2

- Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.
- Witten, D. M., & Tibshirani, R. J. (2019). An Introduction to Statistical Learning: with Applications in R. Springer.
- Tufte, E. R. (2001). The Visual Display of Quantitative Information. Graphics Press.
- Bailer, A. J., & Piegorsch, W. W. (2015). Analyzing Environmental Data. Wiley.
- Unwin, A., Hawkins, G., Hofmann, H., & Siegl, B. (2015). Graphical Data Analysis with R. CRC Press.
- McKinney, W. (2017). Python for Data Analysis. O'Reilly Media.
- Friendly, M., & Meyer, D. (2016). Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data. CRC Press.
- Chatfield, C. (2003). The Analysis of Time Series: An Introduction. CRC Press.
- Chatterjee, S., & Hadi, A. S. (2015). Regression Analysis by Example. Wiley.
- Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.
- Hubert, M., & Van der Veeken, S. (2008). Outlier Detection for High Dimensional Data. The American Statistician.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). Multivariate Data Analysis. Cengage Learning.
- Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics.
- DeSarbo, W. S., Jedidi, K., & Sinha, I. (2001). Customer Value Analysis in a Heterogeneous Market. Springer.
- Gelman, A., & Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
- Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society.
- Cleveland, W. S. (1993). Visualizing Data. Hobart Press.

TEXTE DE DESCRIPTION DU COURS

Welcome to "Data Analysis Mastery: Unveiling Insights & Patterns!" This all-encompassing course is your ultimate guide to the domains of descriptive statistics and exploratory data analysis (EDA). Whether you're on a journey to becoming a data analyst, scientist, researcher, or merely intrigued by the universe of data exploration and interpretation, this course provides the arsenal to dissect, visualize, and derive meaningful insights from vast datasets.

Introduction to the Foundations of Data
Kick off your journey with a dive into the essentials of data types—numeric, categorical—and their scales, from nominal to ratio. Embrace the significance of data representation, cleaning, and preprocessing. Grasp the art of addressing missing data, treating outliers, and mastering data scaling and normalization techniques.

Visualization and Descriptive Statistics Unveiled
Venture into the realm of "Data Visualization," understanding its pivotal role in both descriptive statistics and EDA. Delve into the world of histograms, box plots, scatter plots, and more, using powerful libraries like Matplotlib, Seaborn, and ggplot2. Subsequently, immerse yourself in the core of "Descriptive Statistics". From measures of central tendency—mean, median, mode—to variability aspects such as variance and standard deviation, equip yourself with the tools to quantitatively analyze data. Explore bivariate statistics, cross-tabulations, measures of association, and visualize relationships with scatter plots and heatmaps.

Mastering Data Distributions and Shape
Deepen your understanding of data distributions. Grasp the intricacies of the Normal distribution, the binomial, the Poisson, and more. Harness the power of skewness, kurtosis, and moments to describe data shapes and characteristics. Further, enrich your knowledge with Z-scores, normality tests, and data transformation techniques to handle skewness.

Exploratory Techniques and Advanced Data Analysis
Elevate your skills in "Advanced EDA Techniques". Discover transformation methods like logarithmic and Box-Cox and delve deep into multivariate exploratory tools such as Principal Component Analysis (PCA), Multidimensional Scaling (MDS), and Cluster Analysis. Time series decomposition, Seasonal Decomposition of Time Series (STL), and autocorrelation functions further expand your repertoire.

Descriptive Statistics and EDA in the Real World
Master practical aspects with "Descriptive Statistics in Practice," addressing real-life scenarios in fields like business, economics, and healthcare. Dive into challenges in "Descriptive Statistics in Big Data," exploring distributed computing methods for massive datasets.

Culmination: From Data to Narratives
The course concludes with "Capstone Projects and Case Studies." Embark on your project, navigating through data collection, preparation, in-depth analysis, and interpretation. Synthesize your knowledge, crafting insightful reports and visualizations that narrate compelling data stories.

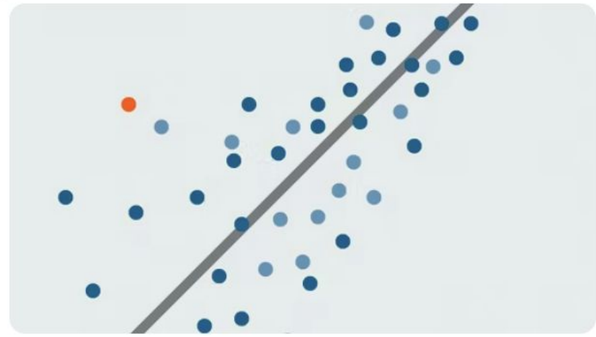
Whether aiming to spearhead data-driven endeavors, refine research skills, or simply unveil the tales data tells, "Data Analysis Mastery: Unveiling Insights & Patterns" is your comprehensive key to the universe of data interpretation and exploration.

DESIGN DE LA CARD

Author: Baptiste Mokas, Weeki

Course Name: Simple Linear Regression

#ExploratoryDataAnalysis
#DataVisualization
#DescriptiveStatistics



Duke University

Linear Regression and Modeling

Compétences que vous acquerrez: Probability & Statistics, Regression, Business Analysis, Data Analysis, General Statistics, Statistical Analysis,…

★ **4.8** (1.7k avis)

Débutant · Course · 1 à 4 semaines