Exploratory Data Analysis 2
  Project Description: 2
  Project Title: 2
  Objective: 2
  Dataset: 2
  Methodology: 3
  Tools and Technologies: 4
  Deliverables: 4
Descriptive Analysis 4
  Project Description: 4
  Project Title: 4
  Objective: 4
  Dataset: 4
  Methodology: 5
  Tools and Technologies: 6
  Deliverables: 6
Diagnostic Analysis 6
    Project Description: Diagnostic Analysis of Sales Decline at XYZ Retail 6
  Project Title: 6
  Objective: 6
  Background: 6
  Dataset: 6
  Methodology: 7
  Tools and Technologies: 7
  Deliverables: 8
  Timeline: 8
Data Wrangling 8
    Project Description: Data Wrangling for Customer Analytics at XYZ Company 8
  Project Title: 8

Objective: 8
Background: 8
Dataset: 8
Methodology: 9
Tools and Technologies: 10
Deliverables: 10
Timeline: 10
Data Quality Control 10
Project Description: Data Quality Control Initiative at ABC Enterprises 10
Project Title: 10
Objective: 10
Background: 11
Scope: 11
Methodology: 11
Deliverables: 12
Timeline: 12
Course Completion Badge 13

Exploratory Data Analysis

Project Description: Exploratory Data Analysis (EDA) on Titanic Dataset

Project Title: Surviving the Titanic: An Exploratory Data Analysis

Objective: The primary goal of this project is to perform an exploratory data analysis (EDA) on the Titanic dataset to uncover patterns, trends, and insights related to passenger survival. By analyzing various features such as age, gender, class, and fare, we aim to understand the factors that influenced the likelihood of survival during the Titanic disaster.

Dataset: The Titanic dataset consists of passenger information from the ill-fated voyage of the RMS Titanic, including details such as:

- Passenger ID: Unique identifier for each passenger
- Survived: Survival status (0 = No, 1 = Yes)

- Pclass: Passenger class (1st, 2nd, 3rd)
- Name: Name of the passenger
- Sex: Gender of the passenger
- Age: Age of the passenger
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard
- Ticket: Ticket number
- Fare: Fare paid by the passenger
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Methodology:

1- Data Collection and Preparation:
    - Load the Titanic dataset using Python libraries like Pandas.
    - Perform initial data cleaning, which includes handling missing values, correcting data types, and renaming columns for clarity.

2- Descriptive Statistics:
    - Generate summary statistics (mean, median, mode) for numerical features (like Age and Fare) and frequency distributions for categorical features (like Pclass and Sex).

3- Data Visualization:
    - Create visualizations to illustrate key insights:
        - Histograms and Boxplots: Analyze the distribution of continuous variables like Age and Fare.
        - Bar Charts: Showcase survival rates across different categories (e.g., Sex, Pclass).
        - Heatmaps: Visualize correlations between numerical variables.

4- Survival Analysis:
    - Compare survival rates:
        - By gender: Determine if there is a significant difference in survival rates between male and female passengers.
        - By class: Analyze how passenger class affected survival chances.
        - By age group: Create age bins to assess survival across different age demographics.

5- Insights and Findings:
    - Summarize the findings based on data visualizations and statistical analyses, highlighting notable trends and patterns (e.g., women and children had higher survival rates, first-class passengers had a significant survival advantage).

6- Conclusion:
   - Discuss the implications of the findings and suggest further analyses or data-driven decisions that could be explored, such as building predictive models to classify survival based on passenger features.

Tools and Technologies:

- Python (Pandas, NumPy, Matplotlib, Seaborn)
- Jupyter Notebook for interactive data exploration
- Any additional data visualization tools like Tableau or Power BI (optional)

Deliverables:

- A comprehensive Jupyter Notebook containing all steps of the analysis, including code, visualizations, and narrative explanations of findings.
- A presentation summarizing key insights and visualizations for stakeholders or peers.

This EDA project not only demonstrates your analytical and programming skills but also highlights your ability to derive meaningful insights from data, making it a valuable addition to your data analyst portfolio.

Descriptive Analysis

Project Description: Descriptive Analysis of Customer Purchase Patterns

Project Title: Understanding Customer Purchase Patterns at XYZ Retail

Objective: The primary goal of this project is to conduct a descriptive analysis of customer purchase data at XYZ Retail. Through this analysis, we aim to summarize key characteristics of customer purchases, identify trends, and generate insights that can inform marketing strategies and inventory management.

Dataset: The dataset includes transactional data from XYZ Retail over the past year, containing the following key features:

- Transaction ID: Unique identifier for each purchase
- Customer ID: Identification number for each customer
- Purchase Date: Date and time of the transaction

- Product Category: Category of the purchased product (e.g., electronics, clothing, groceries)
- Quantity: Number of items purchased
- Price: Total price of the transaction
- Payment Method: Method used for payment (e.g., credit card, cash, digital payment)
- Location: Store location where the purchase was made

Methodology:

1- Data Collection and Preparation:
  - Load the dataset using data analysis tools (e.g., Python, Excel).
  - Perform data cleaning to address missing values, correct data types, and remove duplicates.

2- Descriptive Statistics:
  - Calculate summary statistics for key variables, including:
    - Total sales and average transaction value
    - Number of transactions per month
    - Distribution of purchases by product category
    - Average quantity purchased per transaction

3- Data Visualization:
  - Create visual representations to illustrate findings:
    - Time series graphs showing sales trends over the year.
    - Bar charts displaying the most popular product categories.
    - Pie charts representing the share of different payment methods.
    - Heatmaps of sales by location and time of day.

4- Customer Segmentation:
  - Segment customers based on their purchasing behavior (e.g., high-frequency vs. low-frequency buyers).
  - Analyze the purchasing patterns of different segments.

5- Insights and Findings:
  - Summarize the insights derived from the analysis, highlighting:
    - Peak shopping periods (e.g., holidays, weekends)
    - Trends in product category sales over time
    - Preferences in payment methods across customer segments

6- Recommendations:
  - Provide actionable recommendations based on the findings to inform inventory management, targeted marketing campaigns, and promotional strategies.

Tools and Technologies:

- Python (Pandas, Matplotlib, Seaborn) or Excel for data analysis
- Data visualization tools (Tableau or Power BI) for creating dashboards

Deliverables:

- A detailed report summarizing the methods, findings, and recommendations.
- Visualizations and dashboards to present key insights clearly.
- A presentation for stakeholders to communicate important findings and suggestions for future action.

This descriptive analysis project aims to provide a comprehensive understanding of customer purchase behaviors, enabling XYZ Retail to optimize its operations and enhance customer satisfaction.

Diagnostic Analysis

Project Description: Diagnostic Analysis of Sales Decline at XYZ Retail

Project Title: Investigating the Causes of Sales Decline at XYZ Retail

Objective: The primary goal of this project is to conduct a diagnostic analysis to identify the underlying causes of a recent decline in sales at XYZ Retail. By analyzing various data sources, we aim to uncover the factors contributing to this decline and provide actionable insights for management to formulate effective strategies for improvement.

Background: Over the past six months, XYZ Retail has experienced a noticeable decline in sales, which has prompted management to seek a thorough understanding of the reasons behind this trend. This analysis will not only help in identifying contributing factors but also guide remedial actions to reverse the decline.

Dataset: The analysis will utilize multiple datasets, including:

- Sales Data: Daily sales transactions from the past year, including product categories, transaction amounts, and customer demographics.
- Inventory Data: Records of stock levels for each product category over the same period, indicating any shortages or overstock situations.

- Customer Feedback: Survey data and feedback directly from customers regarding their shopping experiences, preferences, and complaints.
- Market Data: External factors, including regional economic indicators, competitor pricing, and market trends.

Methodology:

1- Data Collection and Preparation:
   - Consolidate and clean datasets from multiple sources to ensure accuracy and consistency.
   - Normalize data to make it suitable for analysis.

2- Trend Analysis:
   - Perform a thorough analysis of sales trends over the last year to identify specific periods and product categories where declines were most significant.

3- Correlation Analysis:
   - Identify correlations between sales decline and other variables, such as inventory levels, customer feedback, and market conditions.
   - Use statistical methods (like regression analysis) to quantify how strongly these factors influence sales.

4- Root Cause Analysis:
   - Conduct focus group discussions or interviews (if feasible) with store staff and management to gather qualitative insights on observed changes in customer behavior and inventory practices.
   - Utilize techniques such as the "5 Whys" or Fishbone Diagram to systematically investigate potential causes.

5- Segmentation Analysis:
   - Segment customers based on behaviors (e.g., frequency of purchase, average transaction size) to analyze differing impacts across segments.

6- Synthesis of Findings:
   - Integrate quantitative and qualitative data to uncover patterns and themes that indicate the most significant factors contributing to the sales decline.

Tools and Technologies:

- Data analysis tools (Python with libraries like Pandas and Scikit-learn, R, or SQL) for metrics calculations and correlation analysis.
- Visualization tools (Tableau or Power BI) for presenting findings and trends clearly to stakeholders.

Deliverables:

- A comprehensive diagnostic report that outlines the analysis process, findings, and confirmed root causes of the sales decline.
- Visualizations and dashboards summarizing key metrics and trends.
- Actionable recommendations for management, focusing on strategies to address identified issues and improve sales performance.

Timeline:

- Expected completion of the project: 6 weeks from project kickoff, including regular check-ins with stakeholders to align findings and recommendations.

This diagnostic analysis aims to provide a clear understanding of the reasons behind the decline in sales at XYZ Retail, enabling management to take targeted actions to boost revenue and enhance customer satisfaction in the future.

Data Wrangling

Project Description: Data Wrangling for Customer Analytics at XYZ Company

Project Title: Data Wrangling for Enhanced Customer Analytics at XYZ Company

Objective: The primary goal of this project is to perform comprehensive data wrangling to prepare a robust dataset for customer analytics at XYZ Company. By cleaning, transforming, and consolidating data from various sources, the project aims to enhance the accuracy and usability of customer data for subsequent analysis and reporting.

Background: XYZ Company has accumulated customer data from multiple channels, including sales transactions, customer service interactions, and marketing campaigns. However, this data is often inconsistent, incomplete, or fragmented, making it challenging to derive meaningful insights. Effective data wrangling will facilitate better decision-making and more targeted marketing strategies.

Dataset: The data wrangling process will involve various datasets, including:

- Sales Data: Transaction records that include customer IDs, purchase amounts, product details, and timestamps.

- Customer Information: Demographic details such as age, gender, location, and account creation date.
- Customer Service Records: Logs of customer inquiries, complaints, and resolutions.
- Marketing Interaction Data: Email and campaign response data, including open rates and click-through rates.

Methodology:

1- Data Collection:
  - Gather datasets from various sources, including internal databases, CRM systems, and third-party marketing platforms.
  - Ensure that all relevant datasets are identified for a comprehensive customer profile.

2- Data Assessment:
  - Conduct an initial assessment of the data quality to identify issues such as missing values, duplicates, and inconsistencies across different datasets.
  - Document data types, formats, and any discrepancies.

3- Data Cleaning:
  - Address missing values through appropriate methods (e.g., imputation or exclusion) based on their significance and context.
  - Remove duplicate records and correct inconsistencies in data formats (e.g., date formats, naming conventions).
  - Normalize categorical variables to ensure consistency across datasets (e.g., standardizing customer status as "active," "inactive," etc.).

4- Data Transformation:
  - Perform data type conversions to ensure that all fields are in suitable formats for analysis (e.g., converting strings to datetime objects).
  - Derive new features that may aid in analytics, such as total purchase amounts, frequency of purchases, or customer tenure.
  - Aggregate data as necessary to ensure that it aligns with the intended analysis (e.g., summarizing monthly sales per customer).

5- Data Consolidation:
  - Merge datasets into a unified customer database, ensuring that all relevant information is linked accurately through unique identifiers (e.g., customer ID).
  - Create a comprehensive view of each customer by combining sales, support, and marketing data.

7- Documentation and Validation:
  - Document the data wrangling process, including data sources, cleaning methods, and transformations applied to the dataset.

    - Validate the final dataset through exploratory data analysis (EDA) to confirm accuracy and completeness.

Tools and Technologies:

- Python (using libraries like Pandas and NumPy) or R for data manipulation and cleaning.
- SQL for data extraction and initial assessment of data from relational databases.
- Jupyter Notebook or RStudio for interactive data wrangling and documentation.
- Visualization tools (like Matplotlib or Seaborn) to assist with EDA and quality checks.

Deliverables:

- A cleaned and transformed customer dataset ready for analysis, available in a suitable format (e.g., CSV, Excel Database).
- A comprehensive report documenting the data wrangling process, including challenges encountered, methods employed, and final dataset characteristics.
- Visualizations illustrating the key data insights and confirmations of data quality checks conducted during the process.

Timeline:

- Expected completion of the project: 6 weeks, including phases for assessment, cleaning, transformation, and documentation.

This data wrangling project aims to establish a high-quality dataset that enables XYZ Company to conduct effective customer analytics, ultimately enhancing marketing strategies, improving customer service, and driving overall business growth.

Data Quality Control

Project Description: Data Quality Control Initiative at ABC Enterprises

Project Title: Implementation of Data Quality Control Measures at ABC Enterprises

Objective: The primary objective of this project is to establish a comprehensive Data Quality Control (DQC) framework at ABC Enterprises. This framework will ensure the accuracy, completeness, consistency, and reliability of the organization's data, enhancing decision-making processes and overall business performance.

Background:
As ABC Enterprises continues to expand its operations and data sources, issues related to data quality have surfaced, including inaccuracies, duplicate records, and inconsistent formats. Poor data quality can lead to misguided business strategies, inefficiencies, and regulatory compliance risks. This project aims to implement robust data quality control measures to mitigate these issues.

Scope:
The project will focus on the following key areas:

- Data Profiling: Analyzing existing datasets to assess quality levels.
- Data Cleansing: Developing processes to correct inaccuracies and eliminate duplicates.
- Data Validation: Implementing validation rules and checks to ensure data integrity.
- Monitoring and Reporting: Establishing ongoing monitoring processes and dashboards to track data quality metrics.
- Training and Awareness: Creating training programs for staff on data quality best practices.

Methodology:

1- Current State Assessment:
  - Conduct a thorough analysis of current data sources, workflows, and existing data quality challenges.
  - Identify the key datasets that significantly impact business operations and decision-making.

2- Data Profiling:
  - Utilize data profiling tools to assess the quality of identified datasets, focusing on completeness, uniqueness, validity, consistency, and accuracy.
  - Document findings to highlight areas requiring immediate attention.

3- Establish Data Quality Metrics:
  - Define clear data quality metrics and key performance indicators (KPIs) to evaluate and track data quality over time, such as error rates, duplicate records, and compliance with data standards.

4- Data Cleansing Processes:
  - Develop and implement procedures for data cleansing, which may include:
    - Removing duplicates and correcting errors.
    - Standardizing data formats and values.
    - Filling in missing values using appropriate imputation techniques.

5- Validation Rules and Procedures:
  - Set up validation rules for new data entries to reduce the risk of poor-quality data being introduced into the system.
  - Create data entry guidelines to promote consistency and accuracy.

6- Monitoring and Reporting:
  - Implement monitoring tools and dashboards that provide real-time data quality metrics and alerts for significant deviations.
  - Schedule regular reports to review data quality trends and performance against established KPIs.

7- Training and Best Practices:
  - Develop training materials and conduct workshops to educate employees on data quality principles, the importance of maintaining data integrity, and procedural best practices.
  - Foster a culture of accountability where employees recognize their role in ensuring data quality.

8- Feedback Mechanism:
  - Establish a feedback loop to continually assess and improve data quality processes based on user input and observed results.

Tools and Technologies:

- Data quality tools (such as Informatica Data Quality, Talend, or Trifacta) for profiling and cleansing.
- Data visualization tools (like Tableau or Power BI) for monitoring and reporting on data quality metrics.
- SQL or Python for data cleansing and automated validation scripts.

Deliverables:

- A comprehensive Data Quality Control plan detailing processes, metrics, and responsibilities.
- Documentation of data quality metrics and KPIs being tracked.
- Cleaned and validated datasets ready for analysis and reporting.
- Training resources, including materials and workshops, are designed to educate staff on data quality practices.
- A monitoring dashboard that visualizes data quality metrics in real-time.

Timeline:

- Expected completion of the project: 8 weeks, including assessment, implementation, training, and monitoring setup.

This Data Quality Control initiative aims to empower ABC Enterprises to enhance its data integrity and reliability, resulting in improved decision-making, operational efficiency, and compliance with regulatory requirements.

Course Completion Badge

Share your course completion badge in your portfolio. You can claim it using a module named “Badges and completion certificates” in your AWS Academy CF course.