IT SKILLS and DATA ANALYSIS II

SKILL ENHANCEMENT COURSE (SEC) SEM III

DATE: 5th November 2023

INSTRUCTIONS:

- Do any two of the five case studies below.
- Please read the whole case study.
- Prepare a comprehensive analysis.
- Summarize everything in a PowerPoint presentation (with regression tables and scatter diagrams).
- You can form a group of 2 or 3 [not more than that].
- Presentations to be held on 25th November 2023 (tentative)
- Submit your ppts on mail ID: maycommail97@gmail.com

Case Study: Predicting House Prices (house_prices_dataset)

Background: You are a data analyst working for a real estate company. Your company is interested in understanding the factors that influence house prices. They have provided you with a dataset containing information about various houses, including features like square footage, number of bedrooms, number of bathrooms, and the sale prices.

Objective: Your goal is to build regression models to predict house prices based on different sets of features.

Steps to Perform:

1. Simple Linear Regression:
    - Choose a feature (e.g., square footage) as the independent variable and sale price as the dependent variable.
    - Create a scatter diagram to visualize the relationship between square footage and sale price.
    - Perform simple linear regression using Excel's regression analysis tool to predict sale price based on square footage.
    - Interpret the results, including the regression equation.
2. Coefficient of Determination and Correlation:
    - Calculate the coefficient of determination (R-squared) to assess how well the regression model fits the data.
    - Calculate the coefficient of correlation (Pearson's r) between square footage and sale price.
3. Multiple Linear Regression:
    - Select multiple features (e.g., square footage, number of bedrooms, number of bathrooms) as independent variables and sale price as the dependent variable.

Mayank

- Perform multiple linear regression using Excel's regression analysis tool to predict sale price based on these features.
- Interpret the results, including the regression equation.

4. Comparison and Evaluation:
   - Compare the results of simple and multiple linear regression models.
   - Evaluate which model provides a better prediction of house prices.
5. Scatter Diagrams:
   - Create scatter diagrams to visualize the relationships between the selected independent variables and the sale price.
6. Simple Hypothesis Testing:
   - Perform a hypothesis test to determine if any of the independent variables have a statistically significant impact on house prices.
   - Interpret the results of the hypothesis test.
7. Conclusion and Recommendations:
   - Summarize your findings and provide recommendations based on the regression models.

Case Study: Analyzing Income Inequality (income_inequality_dataset)

Background: You are a data analyst working for a research organization focused on economic issues. Your organization is interested in understanding the factors that contribute to income inequality within a specific region. They have provided you with a dataset containing information about various socioeconomic variables and income levels of individuals in the region.

Objective: Your goal is to investigate the relationships between different socioeconomic variables and individual income levels to gain insights into income inequality.

Steps to Perform:

1. Simple Linear Regression:
   - Select a variable (e.g., education level) as the independent variable and individual income level as the dependent variable.
   - Create a scatter diagram to visualize the relationship between education level and income.
   - Perform simple linear regression using Excel's regression analysis tool to predict income based on education level.
   - Interpret the results, including the regression equation.

Mayank

2. Coefficient of Determination and Correlation:
    - Calculate the coefficient of determination (R-squared) to assess how well the regression model fits the data.
    - Calculate the coefficient of correlation (Pearson's r) between education level and income.
3. Multiple Linear Regression:
    - Select multiple variables (e.g., education level, years of experience, age) as independent variables and individual income level as the dependent variable.
    - Perform multiple linear regression using Excel's regression analysis tool to predict income based on these features.
    - Interpret the results, including the regression equation.
4. Comparison and Evaluation:
    - Compare the results of simple and multiple linear regression models.
    - Evaluate which model provides a better prediction of individual income levels.
5. Scatter Diagrams:
    - Create scatter diagrams to visualize the relationships between the selected independent variables and individual income levels.
6. Simple Hypothesis Testing:
    - Perform a hypothesis test to determine if any of the independent variables have a statistically significant impact on individual income levels.
    - Interpret the results of the hypothesis test.
7. Conclusion and Recommendations:
    - Summarize your findings and provide recommendations based on the regression models.

Case Study: Analyzing Poverty and Access to Education (poverty_education_dataset)

Background: You are a data analyst working for an NGO focused on addressing social issues, particularly poverty and access to education. Your organization is interested in understanding the factors that contribute to educational attainment in impoverished communities. They have provided you with a dataset containing information about individuals' socioeconomic status, family size, and educational attainment.

Objective: Your goal is to investigate how socioeconomic factors and family size impact educational attainment in this community.

Mayank

Steps to Perform:

1. Data Exploration and Preparation:
    - Load the dataset into Excel.
    - Explore the data to understand its structure and characteristics.
    - Clean and preprocess the data if necessary (e.g., handle missing values, categorical variables, etc.).
2. Simple Linear Regression:
    - Choose a variable (e.g., family income) as the independent variable and educational attainment (e.g., years of schooling) as the dependent variable.
    - Create a scatter diagram to visualize the relationship between family income and educational attainment.
    - Perform simple linear regression using Excel's regression analysis tool to predict educational attainment based on family income.
    - Interpret the results, including the regression equation.
3. Coefficient of Determination and Correlation:
    - Calculate the coefficient of determination (R-squared) to assess how well the regression model fits the data.
    - Calculate the coefficient of correlation (Pearson's r) between family income and educational attainment.
4. Multiple Linear Regression:
    - Select multiple variables (e.g., family income, family size) as independent variables and educational attainment as the dependent variable.
    - Perform multiple linear regression using Excel's regression analysis tool to predict educational attainment based on these features.
    - Interpret the results, including the regression equation.
5. Comparison and Evaluation:
    - Compare the results of simple and multiple linear regression models.
    - Evaluate which model provides a better prediction of educational attainment based on socioeconomic factors and family size.
6. Scatter Diagrams:
    - Create scatter diagrams to visualize the relationships between the selected independent variables (socioeconomic factors, family size) and educational attainment.
7. Simple Hypothesis Testing:
    - Perform a hypothesis test to determine if any of the independent variables have a statistically significant impact on educational attainment.

Mayank

- Interpret the results of the hypothesis test.

8. Conclusion and Recommendations:
    - Summarize your findings and provide recommendations based on the regression models. Consider how interventions related to income support, family planning, and educational resources could potentially improve educational attainment in impoverished communities.

Case Study: Sales Prediction (sales_data)

Background: You work for a retail company that sells electronic gadgets. The company wants to understand the factors that influence the sales of their products. They have collected data on the following variables for the past year:

1. Advertising Spend (in dollars): The amount spent on advertising for each product.
2. Number of Store Visits: The number of people who visited the store where the product was displayed.
3. Online Presence (1 for Yes, 0 for No): Indicates whether the product was available for purchase online.
4. Price (in dollars): The retail price of the product.
5. Sales (in units): The number of units sold.

Objective: Your objective is to build regression models to predict sales based on different combinations of independent variables. Additionally, you need to assess the strength of the relationships and make recommendations for future marketing efforts.

Steps:

1. Scatter Diagrams:
    - Create scatter plots to visualize the relationships between each independent variable and the sales.
2. Simple Linear Regression:
    - Start by performing a simple linear regression using the Advertising Spend as the independent variable and Sales as the dependent variable. Use Excel's regression tool to calculate the regression equation.
3. Coefficient of Determination and Correlation:
    - Calculate the coefficient of determination (R-squared) to measure the proportion of the variance in Sales that can be predicted from Advertising Spend. Also, calculate the coefficient of correlation between the two variables.
4. Interpretation:

Mayank.

    - Interpret the regression equation, R-squared value, and correlation coefficient to explain the relationship between Advertising Spend and Sales.

5. Multiple Linear Regression:
    - Next, perform a multiple linear regression using a combination of independent variables (e.g., Advertising Spend, Number of Store Visits, Online Presence, and Price) to predict Sales.

6. Coefficient of Determination and Correlation (Multiple Regression):
    - Calculate the coefficient of determination (R-squared) for the multiple regression model and correlation coefficients for each independent variable.

7. Model Comparison:
    - Compare the R-squared values and coefficients of correlation for the simple and multiple regression models. Determine which model provides a better fit.

8. Hypothesis Testing:

- Conduct a hypothesis test to determine whether the coefficients of the independent variables in the multiple regression model are statistically significant. Use the t-test or p-value approach.

11. Recommendations:

- Based on your analysis, make recommendations to the company on which factors have the most impact on sales and provide suggestions for optimizing marketing efforts.

12. Visualization:

- Create visualizations (charts, graphs) to present your findings and recommendations.

Remember to document your methodology, assumptions, and interpretations clearly in your report. This case study will give you a comprehensive understanding of regression analysis and its applications in real-world scenarios using Excel.

Case Study: Predicting High School Graduation Rates (social_issues_dataset)

Background: You are a researcher studying the factors that influence high school graduation rates in different communities. You believe that various social and economic factors may play a role in determining graduation rates.

Objective: Your goal is to develop a regression model to predict high school graduation rates based on a set of independent variables.

Data: You have collected data from 1000 different communities. For each community, you have gathered the following information:

Mayank.

1. Average household income (in thousands of dollars)
2. Percentage of single-parent households
3. Average class size
4. Percentage of students receiving free/reduced lunch
5. Crime rate (per 1000 residents)

Procedure:

1. Create a scatter plot for each independent variable against the dependent variable (graduation rate) to visualize their relationships.
2. Simple Linear Regression:
    - Conduct simple linear regression for each independent variable separately against the graduation rate.
    - For example, predict graduation rate using only one independent variable (e.g., average household income).
3. Coefficient of Determination and Correlation:
    - Calculate the coefficient of determination (R-squared) for each simple linear regression model to understand how well each variable predicts graduation rates.
    - Calculate the coefficient of correlation (Pearson's r) between each independent variable and graduation rates.
4. Multiple Linear Regression:
    - Build a multiple linear regression model using all the independent variables (average household income, single-parent households, class size, free/reduced lunch, crime rate) to predict graduation rates.
5. Interpretation:
    - Interpret the coefficients of the multiple linear regression model. Which variables have a significant impact on graduation rates?
6. Hypothesis Testing:
    - Perform hypothesis tests to determine which independent variables are statistically significant in predicting graduation rates. Use a significance level of 0.05.
7. Model Evaluation:
    - Evaluate the overall performance of your multiple regression model using appropriate metrics (e.g., R-squared, adjusted R-squared).
8. Predictions:
    - Use the model to make predictions on graduation rates for new communities.

Mayank

9. Discussion and Conclusion:

   - Summarize your findings. Which variables have the most significant impact on graduation rates? Are there any unexpected results?

DATASETS:

https://drive.google.com/drive/folders/15Vq4znA48YssACuUp7uMpCTCB79EK0-3?usp=sharing

Mayank