CHANAKYA UNIVERSITY

SCHOOL OF MATHEMATICS AND NATURAL SCIENCES

CHANAKYA UNIVERSITY

Term Paper- Statistics for Data Science

ANKIT AJAY MISHRA

Master of Science (Data Science)

Course: Statistics for Data Science

Semester: Second Semester

Instructor: Prof. Naresh Dixit

Submission Date: August 19, 2024

Acknowledgement

I would like to express my deepest gratitude to Prof. Naresh Dixit for his invaluable guidance, encouragement, and support throughout the completion of this term paper on "Statistics for Data Science." His deep knowledge and passion for the subject have been a source of inspiration and have significantly contributed to my understanding of the complex concepts involved in this field.

I am especially thankful to Prof. Dixit for his patience and for taking the time to clarify doubts and provide constructive feedback, which has been instrumental in shaping the direction and quality of this work. His ability to make complex statistical concepts accessible and engaging has greatly enhanced my learning experience.

Finally, I would like to thank the institution for providing a conducive learning environment and access to the resources necessary for the completion of this term paper. The availability of academic materials, software tools, and research facilities has played a crucial role in the successful execution of this project.

This term paper is a product of collective efforts and shared knowledge, and I am sincerely grateful to everyone who has contributed to its realization.

-Ankit Ajay Mishra

INDEX

Dataset link: https://www.data.gov.in/catalog/branchwise-operational-performance-2021-2022

Section A: Descriptive Statistics (20 marks)

1. Define descriptive statistics and explain their significance in data analysis. (5 marks)

Definition: Descriptive statistics are numerical measures that describe the main features of a dataset. They summarize and organize data in an informative way. The most common descriptive statistics include measures of central tendency (mean, median, mode), measures of variability (range, variance, standard deviation), and measures of distribution shape (skewness, kurtosis).

Significance in Data Analysis: Descriptive statistics are essential in data analysis because they provide simple summaries about the sample and the measures. Such statistics form the basis of virtually every quantitative analysis of data, allowing analysts to understand the distribution and variability of the data before performing more complex analyses.

a) Summarization: Condenses large datasets into manageable insights
b) Pattern identification: Reveals trends and relationships in the data
c) Data quality assessment: Helps identify outliers or data entry errors
d) Foundation for further analysis: Provides a basis for inferential statistics
e) Communication: Presents data in an easily understandable format

2. Given the following dataset, calculate the mean, median, mode, variance, and standard deviation: (5 marks)

All of the operations are done using python code and I will attach the code separately. Here I am just writing the theory part

Mean: The average of all values, calculated by summing all values and dividing by the count of values. $\mu = (\Sigma x) / n$, where x are individual values and n is the number of values.

Median: The middle value when the data is sorted. For an odd number of values, it's the middle value. For an even number, it's the average of the two middle values.

Mode: The most frequently occurring value in the dataset.

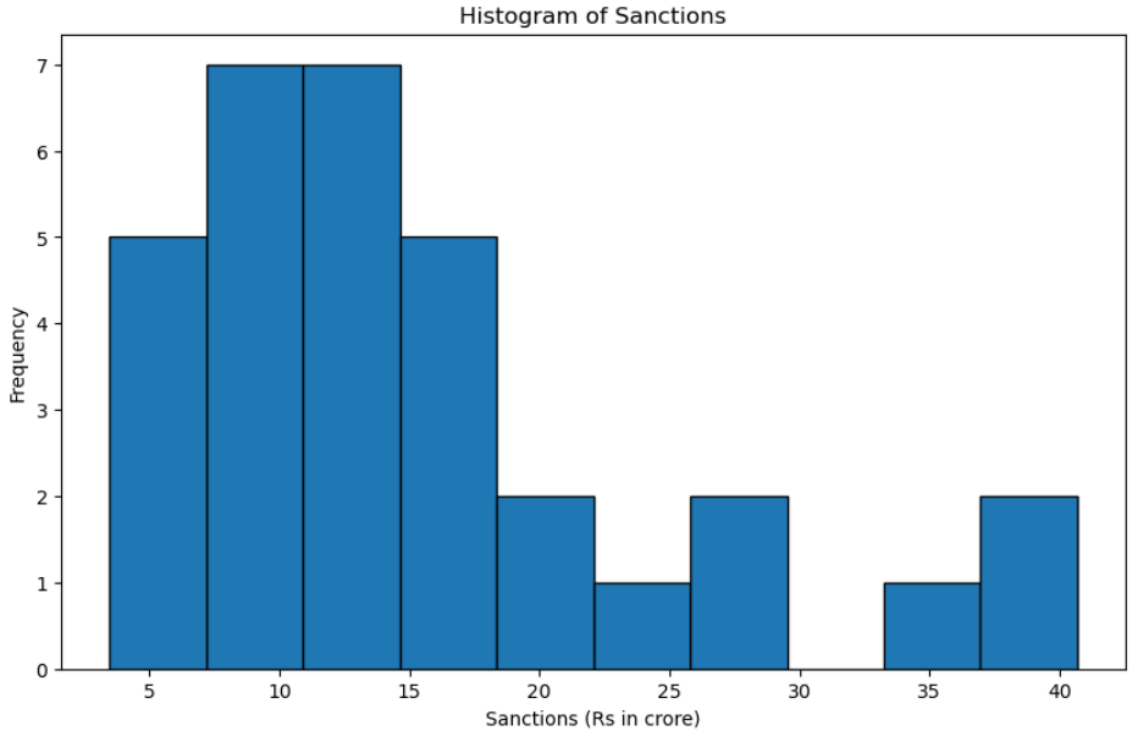
Variance: A measure of variability, calculated as the average squared deviation from the mean. $\sigma^2 = \Sigma(x - \mu)^2 / n$, where x are individual values, μ is the mean, and n is the number of values.

Standard Deviation: The square root of the variance, representing the average deviation from the mean. $\sigma = \sqrt{\sigma^2}$

Mean: 15.41
Median: 13.12
Mode: 11.20
Variance: 91.79
Standard Deviation: 9.58

3. Create a histogram and a box plot for the dataset. Interpret the results. (10 marks)

Histogram:



Box Plot:

Q1: 8.80
Q3: 16.71
IQR: 7.91
Lower bound: -3.06
Upper bound: 28.58
Number of outliers: 4
Outliers: [29.19, 40.69, 35.45, 40.08]

Interpretation of results:

1. Histogram: The histogram displays the frequency distribution of sanctions across different ranges. It helps visualize the shape of the distribution, including any skewness or multimodality.

2. Box plot: The box plot shows the five-number summary (minimum, Q1, median, Q3, maximum) and potential outliers. The box represents the interquartile range (IQR), with the median shown as a line inside the box.

3. Outliers: Data points falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered potential outliers. These are plotted as individual points on the box plot.

4. Skewness: If the median line in the box plot is not centered within the box, it suggests skewness in the data. The histogram shape also indicates skewness.

5. Variability: The width of the box in the box plot and the spread of the histogram indicate the variability in the sanctions data.

6. Central tendency: The peak of the histogram and the median line in the box plot provide insights into the central tendency of the sanctions data.

Section B: Inferential Statistics (25 marks)

1. Explain the concept of inferential statistics. (5 marks)

   Inferential statistics is a branch of statistics that uses sample data to make predictions or inferences about a larger population. It allows us to draw conclusions beyond the immediate data alone. The main concepts in inferential statistics include:

   a) Population vs. Sample: The population is the entire group being studied, while a sample is a subset of that population.

   b) Parameter vs. Statistic: A parameter is a characteristic of the population, while a statistic is a characteristic of the sample.

   c) Probability Distributions: These describe the likelihood of different outcomes in a population.

   d) Hypothesis Testing: A method for making decisions about population parameters based on sample data.

   e) Confidence Intervals: A range of values that is likely to contain the true population parameter with a certain level of confidence.

   Inferential statistics is crucial because it allows us to make educated guesses about large populations based on smaller, more manageable samples. This is particularly useful when studying entire populations is impractical or impossible.

2. Hypothesis Testing: (15 marks)

   Let's perform a hypothesis test to determine if there's a significant difference in sanctions between two groups of offices. I have split the offices into two groups based on their median sanction amount and perform an independent t-test.

   [I have written a Python Code for this and have attached the code separately too]

   Null Hypothesis (H0): There is no significant difference in mean sanctions between the two groups of offices.
   Alternative Hypothesis (H1): There is a significant difference in mean sanctions between the two groups of offices.

   t-statistic: -5.429001550540439
   p-value: 6.937424597035295e-06

   Reject the null hypothesis.
   There is a significant difference in mean sanctions between the two groups of offices.

Interpretation of results:

- If the p-value is less than the chosen significance level (usually 0.05), we reject the null hypothesis. This suggests that there is a statistically significant difference in mean sanctions between the two groups of offices.
- If the p-value is greater than or equal to the significance level, we fail to reject the null hypothesis. This suggests that there is not enough evidence to conclude a significant difference in mean sanctions between the two groups.
  The t-statistic represents the difference between the two group means in units of standard error. A larger absolute value of the t-statistic indicates a greater difference between the groups.

3. Confidence Intervals: Calculate a 95% confidence interval for the meaning of a specific feature in the dataset. (5 marks)

   [I have written a Python Code for this and have attached the code separately too]

   Sample Mean: 15.41
   95% Confidence Interval: (11.90, 18.92)

   The confidence interval is calculated using the formula:
   $CI = \bar{X} \pm (t * (s / \sqrt{n}))$
   Where:

- $\bar{X}$ is the sample mean
- t is the t-value from the t-distribution for the given confidence level and degrees of freedom
- s is the sample standard deviation
- n is the sample size
-

Interpretation:

We can be 95% confident that the true population mean of sanctions falls within this interval. This means that if we were to repeat this sampling process many times and calculate the confidence interval each time, about 95% of these intervals would contain the true population mean.
The width of the confidence interval gives us an idea of the precision of our estimate. A narrower interval indicates a more precise estimate, while a wider interval suggests more uncertainty.
This confidence interval provides valuable information about the likely range of the true population mean sanctions, taking into account the variability in our sample data.

Section C: Regression Analysis (25 marks)

1. Define linear regression and its assumptions. (5 marks)

   Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables.

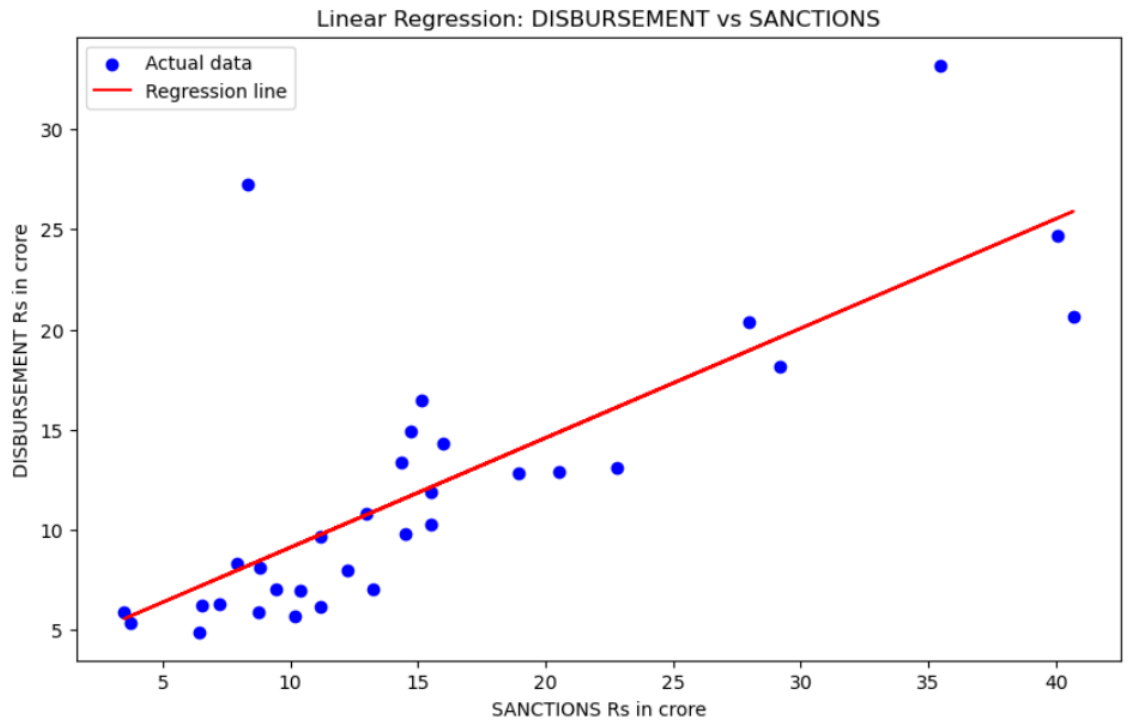
   The basic form of a simple linear regression is:
   $\mathbf{Y = \beta_0 + \beta_1 X + \varepsilon}$

   Where: Y is the dependent variable X is the independent variable β_0 is the y-intercept β_1 is the slope ε is the error term

   Assumptions of linear regression:
   a) Linearity: The relationship between X and Y is linear.
   b) Independence: Observations are independent of each other.
   c) Homoscedasticity: The variance of residual is the same for any value of X.
   d) Normality: For any fixed value of X, Y is normally distributed.
   e) No or little multicollinearity: The independent variables are not highly correlated with each other.

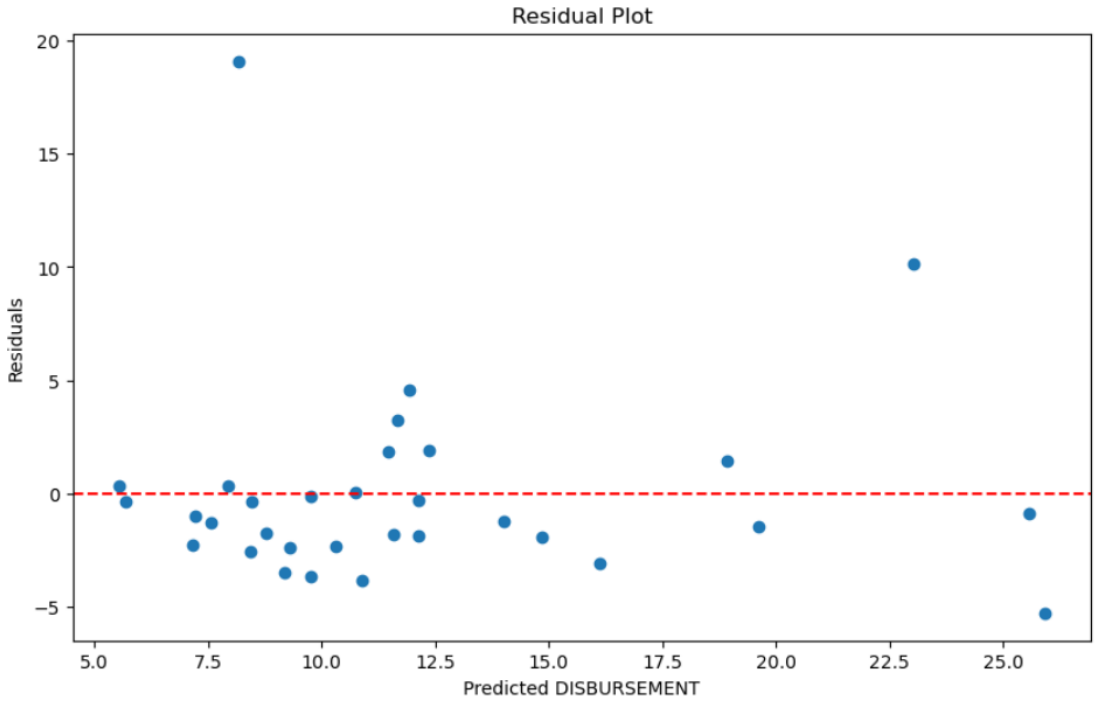
2. Regression Problem:

○ Consider predicting house prices based on features like square footage, number of bedrooms, etc.
○ Build a linear regression model using the dataset.
○ Evaluate the model (R-squared, residuals) and interpret the coefficients. (20 marks)



Linear Regression Results:

Intercept (β_0): 3.6431
Coefficient (β_1): 0.5471
R-squared: 0.5828
Mean Squared Error: 19.6690



Interpretation of results:

1. R-squared: This value represents the proportion of variance in the dependent variable (DISBURSEMENT) that is predictable from the independent variable (SANCTIONS). A higher R-squared indicates a better fit.
2. Coefficients:
   - The intercept (β_0) represents the expected DISBURSEMENT when SANCTIONS is zero.
   - The coefficient (β_1) represents the change in DISBURSEMENT for a one-unit change in SANCTIONS.
3. Residual plot: This helps check the assumptions of homoscedasticity and linearity. If the residuals are randomly scattered around the horizontal line at 0, it suggests these assumptions are met.
4. Scatter plot: This visualizes the relationship between SANCTIONS and DISBURSEMENT, showing how well the linear model fits the data.

Evaluation of assumptions:

1. Linearity: Check if the scatter plot shows a roughly linear relationship.
2. Independence: Assumed based on data collection method.
3. Homoscedasticity: Check if the residual plot shows a consistent spread of residuals.
4. Normality: Can be further checked with a Q-Q plot or histogram of residuals (not included in this code).
5. No multicollinearity: Not applicable for simple linear regression with one predictor.

Section D: Graphs and Visualization (15 marks)

1. Discuss the importance of data visualization

Data visualization is crucial in data analysis for several reasons:

a) Simplifies complex information: Visuals can convey complex patterns and trends in data more efficiently than raw numbers.

b) Facilitates pattern recognition: Humans are inherently visual creatures, and we can quickly identify patterns, trends, and outliers in graphical representations.

c) Enhances communication: Visualizations make it easier to communicate findings to both technical and non-technical audiences.

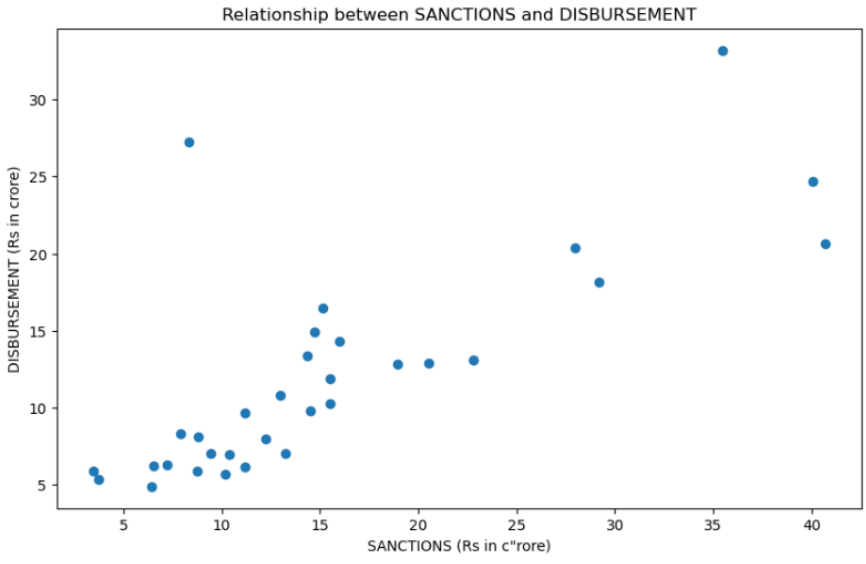
d) Supports decision-making: Well-designed visualizations can provide clear insights that inform better decision-making.

e) Enables comparison: Visualizations allow for easy comparison between different variables or datasets.

f) Highlights relationships: Certain types of plots can reveal relationships between variables that might not be apparent in tabular data.

g) Identifies outliers and anomalies: Visual representations can quickly highlight data points that don't fit the overall pattern.

2. Using Matplotlib or Seaborn, create:
   ○ A scatter plot showing the relationship between two relevant features.
   ○ A bar chart displaying a summary statistic (e.g., mean) for a categorical variable. (10 marks)



3. Explain the insights gained from the visualizations. (5 marks)

Scatter plot insights:

1. Relationship: The scatter plot shows the relationship between SANCTIONS and DISBURSEMENT. If there's a positive slope, it indicates that as SANCTIONS increase, DISBURSEMENT tends to increase as well.
2. Correlation strength: The tightness of the points around an imaginary line indicates the strength of the correlation. A tighter grouping suggests a stronger correlation.
3. Outliers: Any points that are far from the main cluster might represent offices with unusual SANCTIONS-DISBURSEMENT relationships.
4. Range: The plot gives us an idea of the range of both SANCTIONS and DISBURSEMENT across offices.

Bar chart insights:

1. Top performers: The bar chart quickly shows which offices have the highest RECOVERY amounts.
2. Distribution: We can see how RECOVERY amounts vary across the top-performing offices.
3. Comparison: It's easy to compare RECOVERY amounts between different offices visually.
4. Outliers: Any exceptionally high or low bars stand out, potentially indicating offices with unusual performance.

Additional insights:

1. The correlation coefficient quantifies the strength and direction of the relationship between SANCTIONS and DISBURSEMENT. A value close to 1 indicates a strong positive correlation, while a value close to 0 indicates a weak correlation.
2. The mean RECOVERY provides a benchmark against which individual office performance can be compared.

These visualizations and statistics provide a quick and intuitive understanding of the relationships between key variables (SANCTIONS and DISBURSEMENT) and the performance of offices in terms of RECOVERY. They help identify top-performing offices, potential outliers, and overall trends in the data, which can be valuable for decision-making and further analysis.

Section E: Error Calculation and Prediction Model (15 marks)

1. Error Metrics:
○ Define Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
○ Calculate these errors for your regression model. (10 marks)

Definitions:

Mean Absolute Error (MAE): MAE is the average of the absolute differences between predicted values and actual values. It gives an idea of how far the predictions are from the actual values, on average.

Mathematically: MAE = (1/n) * Σ|y_i - ŷ_i| Where y_i are actual values, ŷ_i are predicted values, and n is the number of observations.

Root Mean Squared Error (RMSE): RMSE is the square root of the average of squared differences between predicted values and actual values. It gives more weight to larger errors and is always larger than or equal to MAE.

Mathematically: RMSE = sqrt((1/n) * Σ(y_i - ŷ_i)²)

Error Calculation:

Mean Absolute Error: 4.0880
Root Mean Squared Error: 7.6991
R-squared: 0.0944

[Python code for the same has been provided.]

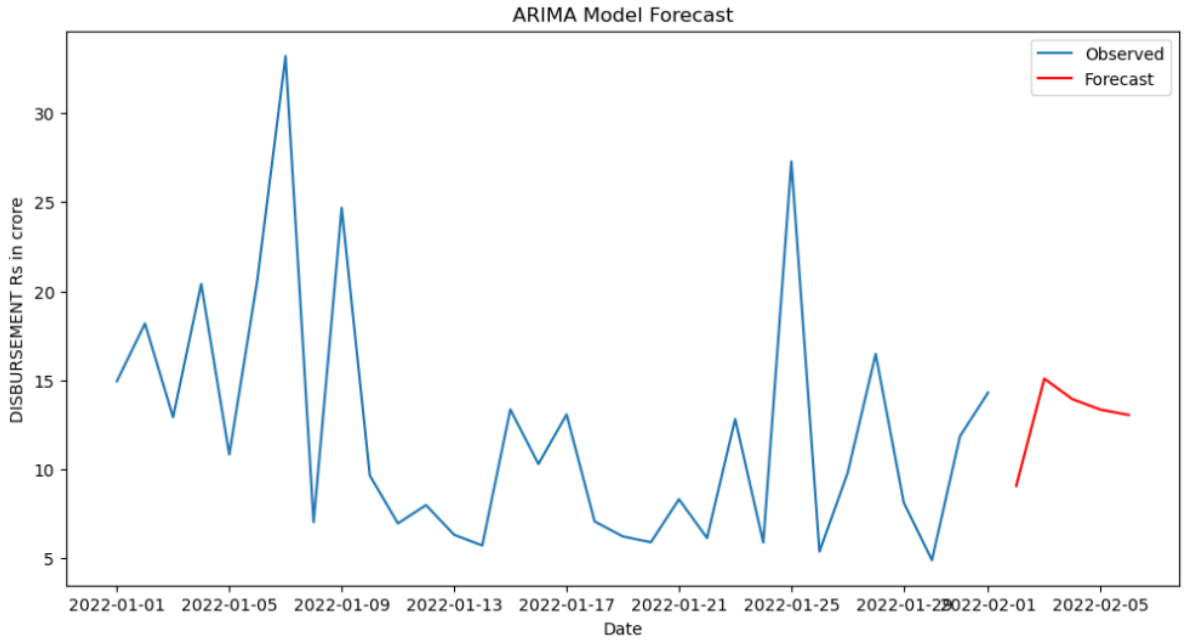
Interpretation:

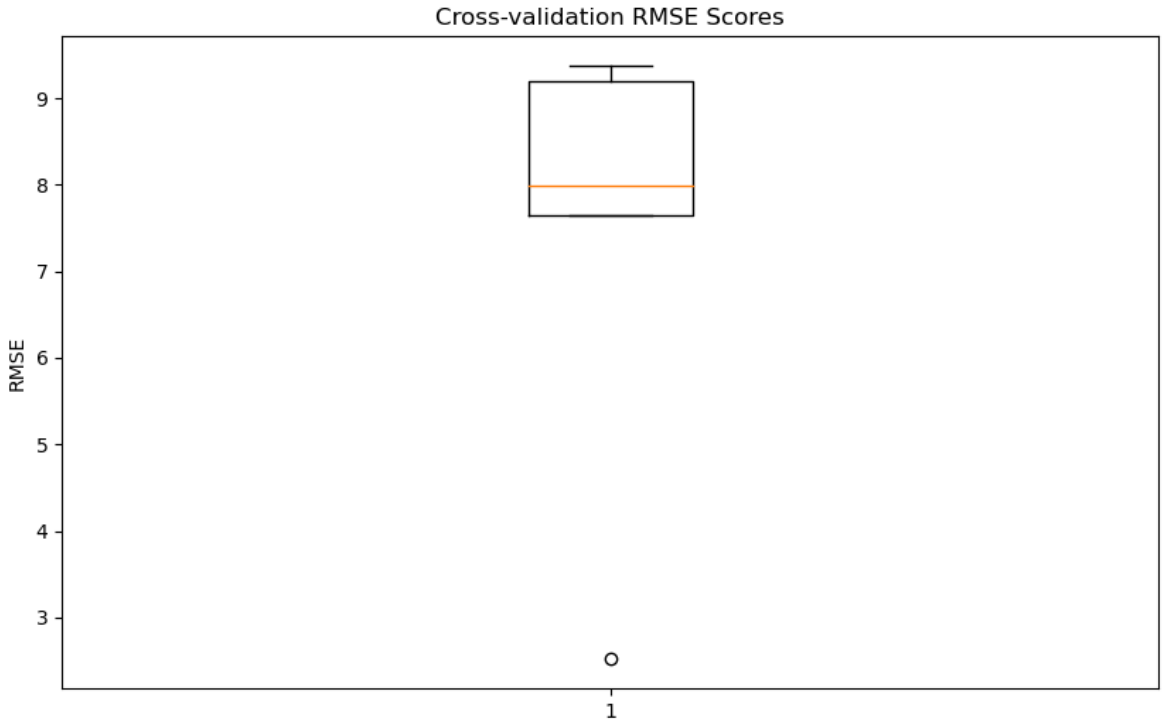
- MAE represents the average absolute difference between predicted and actual DISBURSEMENT values.
- RMSE represents the standard deviation of the residuals (prediction errors). It gives a relatively high weight to large errors.
- A lower value for both MAE and RMSE indicates better model performance.
- R-squared is provided for comparison, showing the proportion of variance in the dependent variable that is predictable from the independent variable.

2. Prediction Model:
○ Choose an appropriate prediction model (e.g., ARIMA, ARMA, SARIMA).
○ Evaluate its performance (box plots) using cross-validation. (5 marks)

For time series prediction, I have use the ARIMA model.

ARIMA(0, 0, 0) RMSE=6.540
ARIMA(0, 0, 1) RMSE=6.709
ARIMA(0, 0, 2) RMSE=7.165
ARIMA(0, 1, 0) RMSE=10.592
ARIMA(0, 1, 1) RMSE=7.179
ARIMA(0, 1, 2) RMSE=7.270
ARIMA(1, 0, 0) RMSE=6.826
ARIMA(1, 0, 1) RMSE=7.093
ARIMA(1, 0, 2) RMSE=6.077
ARIMA(1, 1, 0) RMSE=8.389
ARIMA(1, 1, 1) RMSE=7.350
ARIMA(1, 1, 2) RMSE=7.780
ARIMA(2, 0, 0) RMSE=7.050
ARIMA(2, 0, 1) RMSE=6.842
ARIMA(2, 0, 2) RMSE=6.788
ARIMA(2, 1, 0) RMSE=7.311
ARIMA(2, 1, 1) RMSE=7.553
ARIMA(2, 1, 2) RMSE=7.811
Best ARIMA(1, 0, 2) RMSE=6.077





Cross-validation RMSE scores: [9.371551008717832, 9.194635609261498, 7.646586965972368, 7.994030967920235, 2.5191961901591755]
Mean CV RMSE: 7.345

Interpretation:

1. ARIMA Model:
   - The best ARIMA order (p,d,q) is determined by evaluating different combinations and choosing the one with the lowest RMSE.
   - The forecast plot shows the observed data and the predicted values for the next 5 periods.
   -
2. Cross-validation:
   - The box plot shows the distribution of RMSE scores across different folds of the data.
   - The mean CV RMSE gives an idea of the model's average performance across different subsets of the data.

This analysis provides insights into the ARIMA model's performance in forecasting DISBURSEMENT, its consistency across different subsets of the data, and visualizes both the forecast and the model's performance variability.

Section F: Application-Level Questions (10 marks)

1. Relate statistical concepts to real-world scenarios:

a) Regression Analysis in Real Estate: Real estate agents use regression analysis to predict house prices. They might use features like square footage, number of bedrooms, location, and age of the house to create a model that estimates property values. This helps in pricing homes accurately and understanding which factors most significantly impact property values.

b) Hypothesis Testing in Medicine: Pharmaceutical companies use hypothesis testing when developing new drugs. They might form a null hypothesis that "the new drug has no effect" and an alternative hypothesis that "the new drug improves patient outcomes." Through clinical trials and statistical analysis, they can determine if there's enough evidence to reject the null hypothesis and conclude that the drug is effective.

c) Confidence Intervals in Political Polling: When reporting poll results, news organizations often provide a margin of error, which is related to confidence intervals. For example, they might report that a candidate has 52% support with a margin of error of ±3%. This means they're confident (usually 95% confident) that the true population support lies between 49% and 55%.

d) Time Series Analysis in Stock Market Prediction: Financial analysts use time series models like ARIMA to forecast stock prices or market indices. By analyzing historical data and identifying patterns, they can make predictions about future market behavior, which informs investment strategies.

e) Cluster Analysis in Marketing: Companies use cluster analysis to segment their customer base. By grouping customers based on similarities in purchasing behavior, demographics, or preferences, they can tailor marketing strategies and product offerings to specific customer segments.

2. Discuss the challenges and ethical considerations in applying statistical models to practical problems.

Challenges:

a) Data Quality and Availability: Obtaining high-quality, relevant data can be challenging. Missing data, outliers, or biased samples can lead to inaccurate models and unreliable conclusions.

b) Model Complexity vs. Interpretability: More complex models might provide better predictions but are often harder to interpret. This can be problematic when decisions based on the model need to be explained to stakeholders or the public.

c) Overfitting: Models that fit the training data too closely may perform poorly on new, unseen data. Balancing model complexity with generalizability is an ongoing challenge.

d) Changing Environments: In many real-world scenarios, the underlying relationships in the data can change over time (concept drift), requiring models to be regularly updated or rebuilt.

Ethical Considerations:

a) Bias and Fairness: Statistical models can perpetuate or amplify existing biases in the data. For example, a hiring algorithm trained on historical data might discriminate against certain groups if past hiring practices were biased.

b) Privacy Concerns: The use of personal data in statistical models raises privacy issues. There's a need to balance the benefits of data-driven insights with individuals' rights to privacy and data protection.

c) Transparency and Explainability: Especially in high-stakes decisions (e.g., criminal justice, loan approvals), there's an ethical obligation to provide clear explanations of how statistical models arrive at their conclusions.

d) Misuse or Misinterpretation of Results: Statistical results can be misinterpreted or deliberately misused to support predetermined conclusions. There's an ethical responsibility to present results accurately and in context.

e) Accountability: When decisions are made based on statistical models, it can be unclear who is responsible for negative outcomes – the model creators, the data providers, or the decision-makers using the model.

f) Informed Consent: In many cases, individuals may not be aware that their data is being used in statistical models. There are ethical considerations around obtaining informed consent for data use.

ANNEXURES

The dataset and the Python code for all the sections are available on any of the following links.

GitHub: https://github.com/ankitt02/Statistic_Python

OneDrive: https://chanakyauniversity-my.sharepoint.com/:f:/g/personal/ankitm_msc23_chanakyauniversity_edu_in/EgmkR_6fs2VLlePvBbkU56MBEqGARds9yGtNv98wC93exA?e=vPRayS