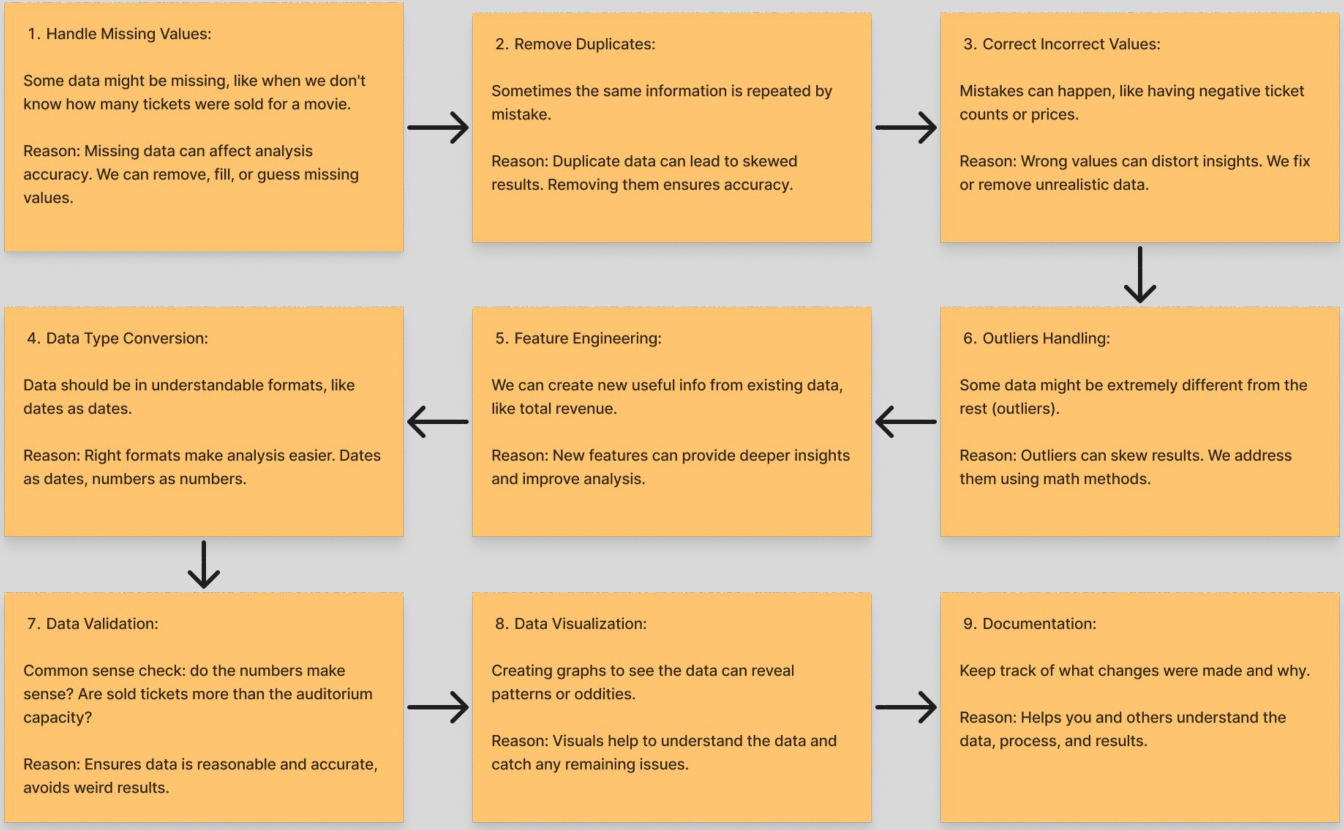
Latihan 1

Cleaning Movie Ticket Sales Dataset



1. Handle Missing Values:

Some data might be missing, like when we don't know how many tickets were sold for a movie.

Reason: Missing data can affect analysis accuracy. We can remove, fill, or guess missing values.

2. Remove Duplicates:

Sometimes the same information is repeated by mistake.

Reason: Duplicate data can lead to skewed results. Removing them ensures accuracy.

3. Correct Incorrect Values:

Mistakes can happen, like having negative ticket counts or prices.

Reason: Wrong values can distort insights. We fix or remove unrealistic data.

4. Data Type Conversion:

Data should be in understandable formats, like dates as dates.

Reason: Right formats make analysis easier. Dates as dates, numbers as numbers.

5. Feature Engineering:

We can create new useful info from existing data, like total revenue.

Reason: New features can provide deeper insights and improve analysis.

6. Outliers Handling:

Some data might be extremely different from the rest (outliers).

Reason: Outliers can skew results. We address them using math methods.

7. Data Validation:

Common sense check: do the numbers make sense? Are sold tickets more than the auditorium capacity?

Reason: Ensures data is reasonable and accurate, avoids weird results.

8. Data Visualization:

Creating graphs to see the data can reveal patterns or oddities.

Reason: Visuals help to understand the data and catch any remaining issues.

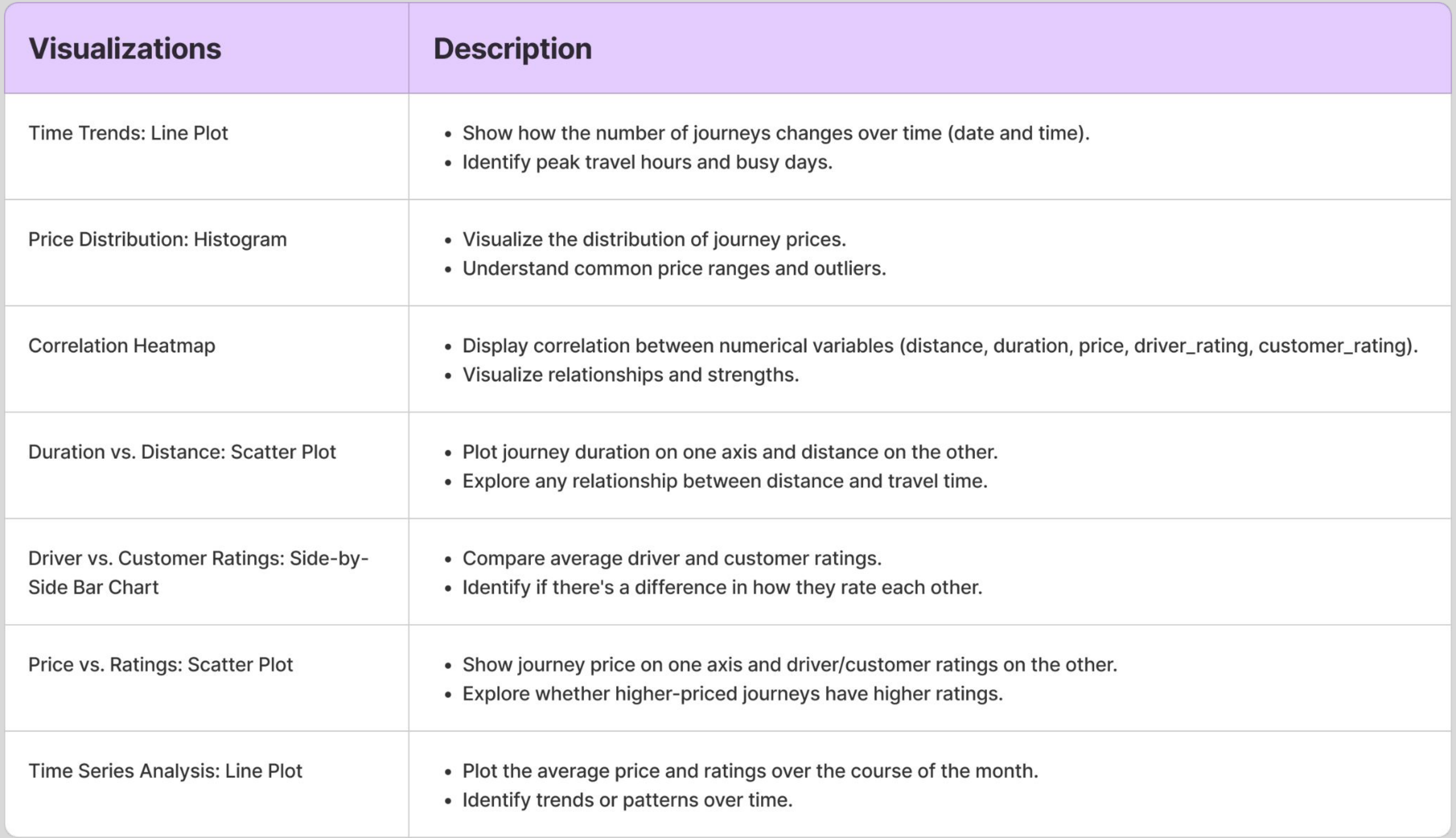
9. Documentation:

Keep track of what changes were made and why.

Reason: Helps you and others understand the data, process, and results.

Latihan 2

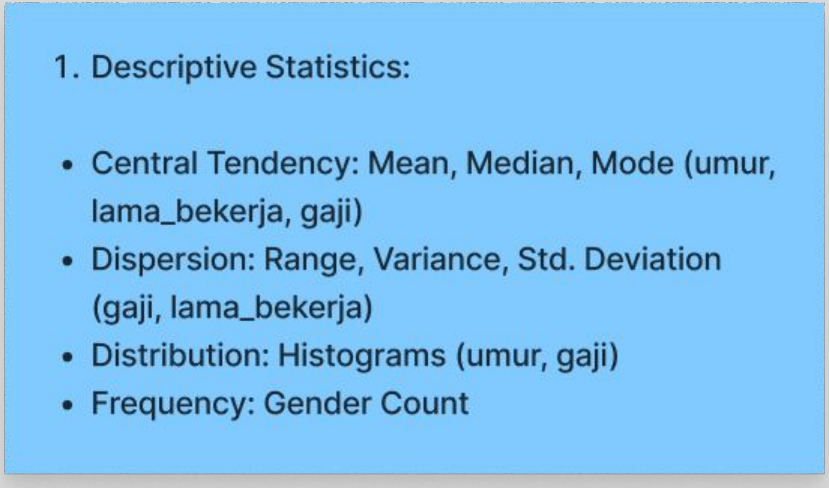
Cleaning Movie Ticket Sales Dataset



Visualizations	Description
Time Trends: Line Plot	• Show how the number of journeys changes over time (date and time). • Identify peak travel hours and busy days.
Price Distribution: Histogram	• Visualize the distribution of journey prices. • Understand common price ranges and outliers.
Correlation Heatmap	• Display correlation between numerical variables (distance, duration, price, driver_rating, customer_rating). • Visualize relationships and strengths.
Duration vs. Distance: Scatter Plot	• Plot journey duration on one axis and distance on the other. • Explore any relationship between distance and travel time.
Driver vs. Customer Ratings: Side-by-Side Bar Chart	• Compare average driver and customer ratings. • Identify if there's a difference in how they rate each other.
Price vs. Ratings: Scatter Plot	• Show journey price on one axis and driver/customer ratings on the other. • Explore whether higher-priced journeys have higher ratings.
Time Series Analysis: Line Plot	• Plot the average price and ratings over the course of the month. • Identify trends or patterns over time.

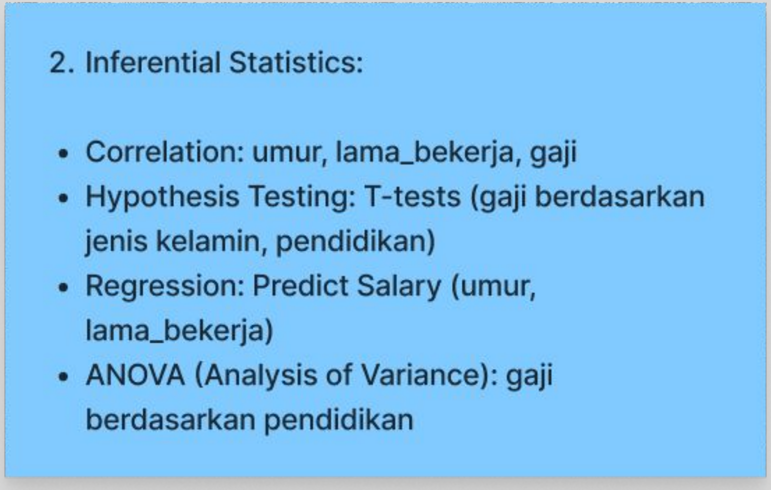
Latihan 3

Analyzing Employee Dataset



1. Descriptive Statistics:

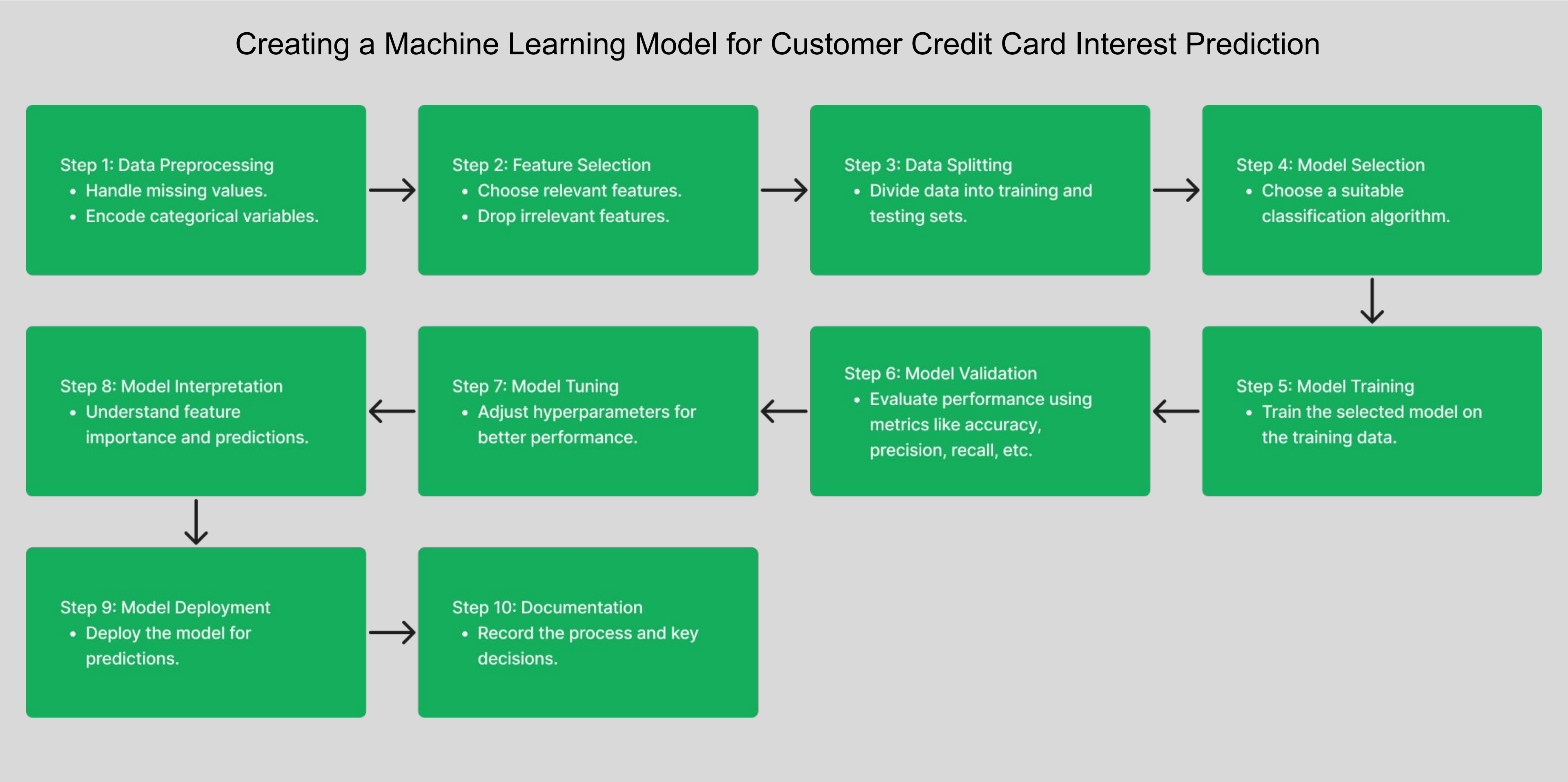
- Central Tendency: Mean, Median, Mode (umur, lama_bekerja, gaji)
- Dispersion: Range, Variance, Std. Deviation (gaji, lama_bekerja)
- Distribution: Histograms (umur, gaji)
- Frequency: Gender Count



2. Inferential Statistics:

- Correlation: umur, lama_bekerja, gaji
- Hypothesis Testing: T-tests (gaji berdasarkan jenis kelamin, pendidikan)
- Regression: Predict Salary (umur, lama_bekerja)
- ANOVA (Analysis of Variance): gaji berdasarkan pendidikan

Latihan 4



Creating a Machine Learning Model for Customer Credit Card Interest Prediction

Step 1: Data Preprocessing
- Handle missing values.
- Encode categorical variables.

Step 2: Feature Selection
- Choose relevant features.
- Drop irrelevant features.

Step 3: Data Splitting
- Divide data into training and testing sets.

Step 4: Model Selection
- Choose a suitable classification algorithm.

Step 5: Model Training
- Train the selected model on the training data.

Step 6: Model Validation
- Evaluate performance using metrics like accuracy, precision, recall, etc.

Step 7: Model Tuning
- Adjust hyperparameters for better performance.

Step 8: Model Interpretation
- Understand feature importance and predictions.

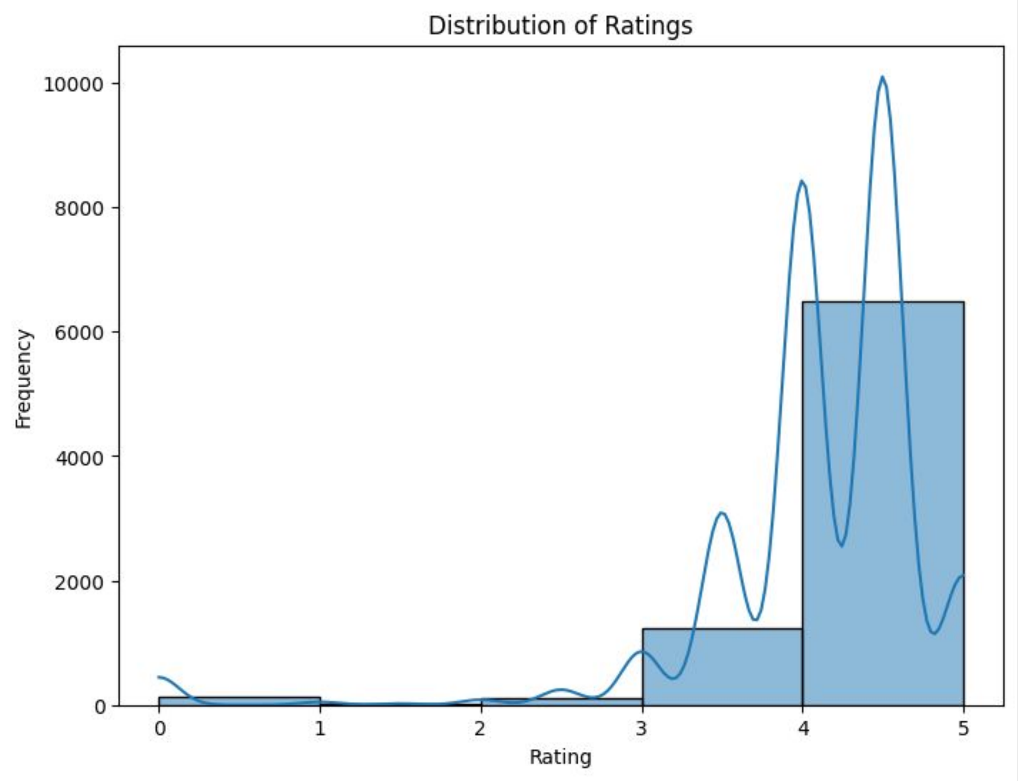
Step 9: Model Deployment
- Deploy the model for predictions.

Step 10: Documentation
- Record the process and key decisions.

Latihan 5

Machine Learning Models and Data Visualization

Data Visualization

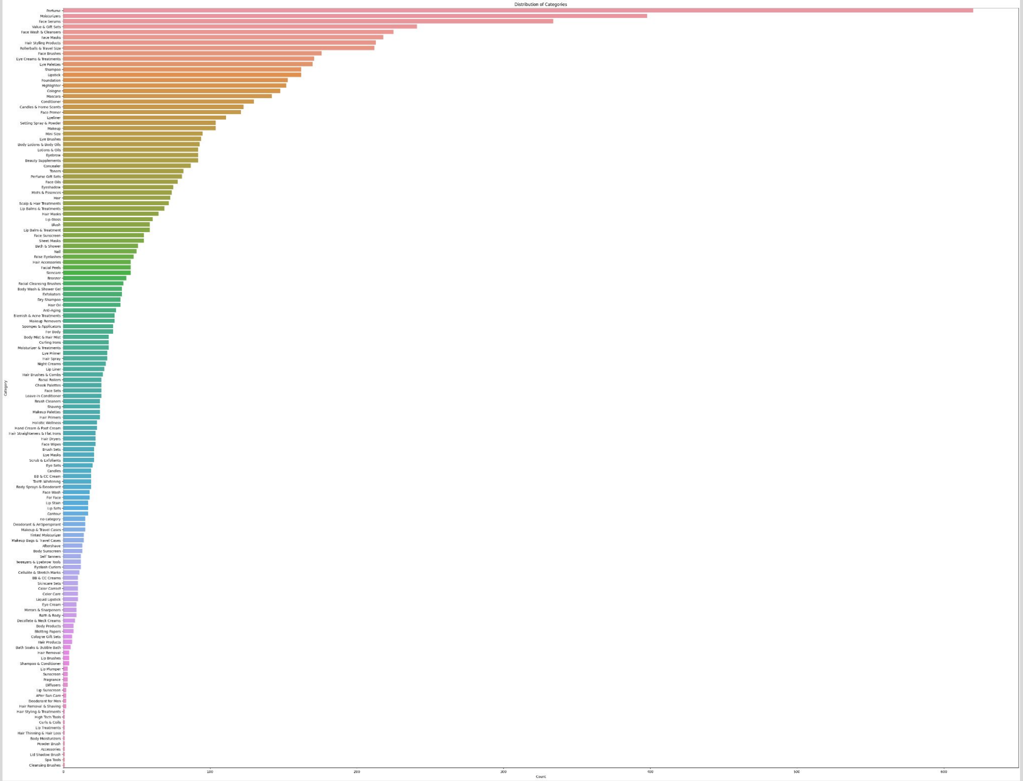


- Peaks or clusters in the histogram indicate the most common rating values, helping to identify trends in customer sentiment.
- This visualization offers insights into whether customers tend to rate products more positively (higher ratings) or negatively (lower ratings) and whether any specific rating values dominate.

Latihan 5

Machine Learning Models and Data Visualization

Data Visualization



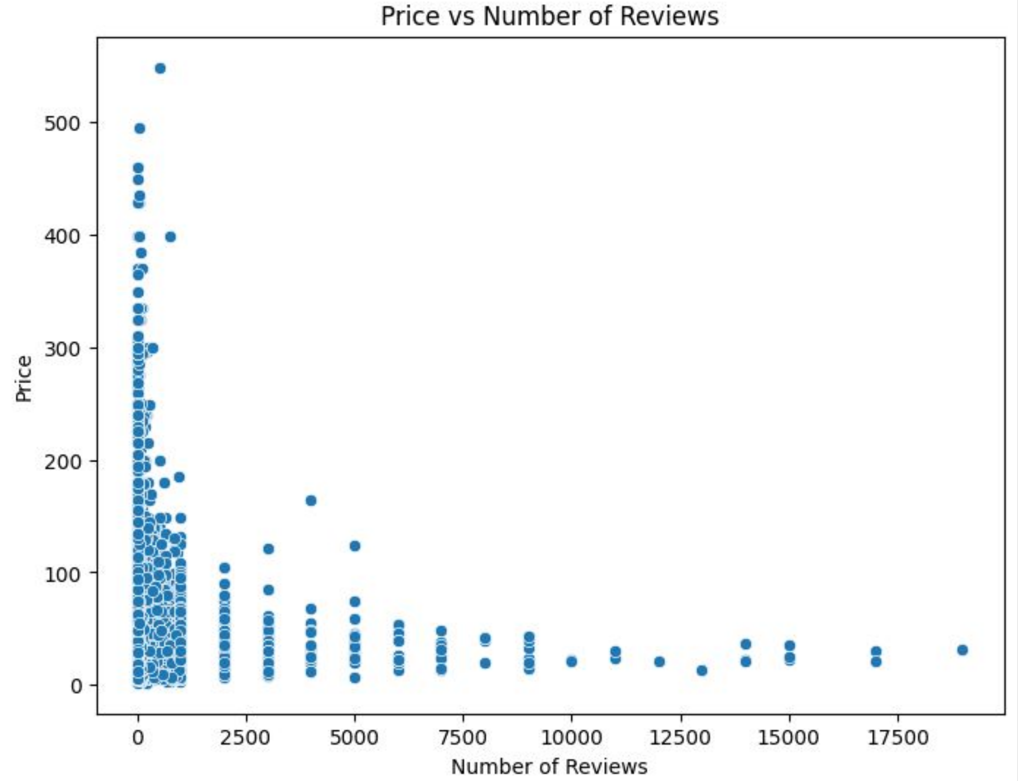
For bigger resolution of the picture

- The count plot displays the frequency of products in each category, providing a visual summary of category popularity.
- The order of categories can provide insights into which categories are most common or least common.
- This visualization assists in identifying which categories have more product offerings, informing marketing strategies and inventory management.

Latihan 5

Machine Learning Models and Data Visualization

Data Visualization

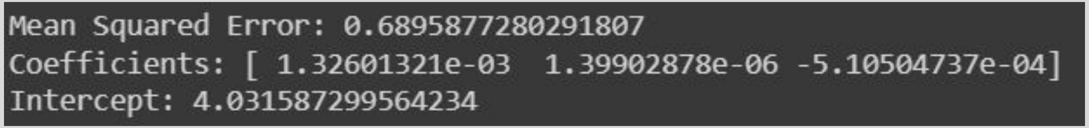


- The scatter plot illustrates the relationship between two numeric variables: 'price' and 'number_of_reviews'.
- Patterns or trends in the scatter plot reveal potential connections between price and customer engagement (reviews).
- The plot may show whether higher-priced products tend to attract more or fewer reviews, helping to understand the impact of pricing on customer interactions.

Latihan 5

Machine Learning Models and Data Visualization

Machine Learning Models



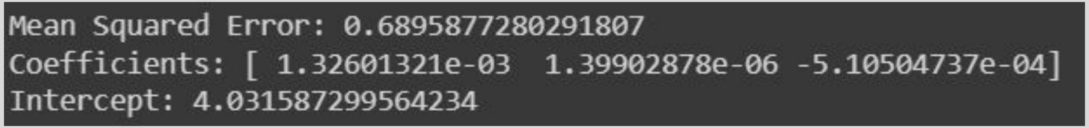
```
Mean Squared Error: 0.6895877280291807
Coefficients: [ 1.32601321e-03  1.39902878e-06 -5.10504737e-04]
Intercept: 4.031587299564234
```

- Mean Squared Error (MSE): The calculated MSE is approximately 0.6896. This means that, on average, the squared difference between the predicted ratings and the actual ratings is around 0.6896. Lower MSE values suggest better predictive accuracy, while higher values suggest greater prediction errors.

Latihan 5

Machine Learning Models and Data Visualization

Machine Learning Models



```
Mean Squared Error: 0.6895877280291807
Coefficients: [ 1.32601321e-03  1.39902878e-06 -5.10504737e-04]
Intercept: 4.031587299564234
```

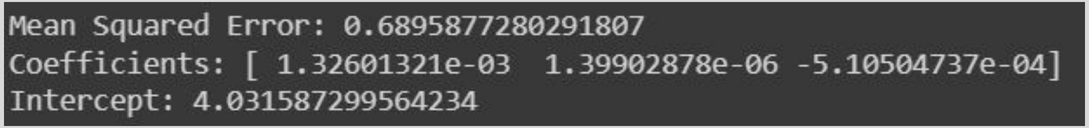
- Coefficients: Coefficients are values assigned to each feature in the model that indicate the strength and direction of their influence on the predicted target variable (in this case, ratings). The coefficients show how a unit change in each feature affects the predicted rating while keeping other features constant.

- The coefficient for 'price' is approximately 0.00133. This means that for every unit increase in the 'price' of a product, the model predicts an increase of about 0.00133 in the product's rating.
- The coefficient for 'love' is approximately 0.0000014. This suggests that for every additional interaction denoted as 'love', the predicted rating increases by around 0.0000014.
- The coefficient for 'value_price' is approximately -0.00051. This indicates that for each unit increase in the 'value_price', the model predicts a decrease of about 0.00051 in the rating.

Latihan 5

Machine Learning Models and Data Visualization

Machine Learning Models



```
Mean Squared Error: 0.6895877280291807
Coefficients: [ 1.32601321e-03  1.39902878e-06 -5.10504737e-04]
Intercept: 4.031587299564234
```

- Intercept: The intercept is the predicted rating when all features are zero. In this case, it's approximately 4.0316. This value provides the baseline predicted rating that the model starts with when all other features are absent or have no influence. It's the y-intercept of the linear regression line.

Latihan 5

Machine Learning Models and Data Visualization

Source Code:

https://github.com/andrewsihotang/pre_test/blob/main/exercise5.ipynb