Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Analysis of Categorical Variables

In our exploration of the categorical columns, several insights emerge:

Seasonal Trends: Bookings peak during the fall season, suggesting a preference for biking during this time. Moreover, there's a notable increase in bookings across all seasons from 2018 to 2019, indicating a positive trend in demand over time.

Monthly Patterns: Months like May, June, July, August, September, and October witness the highest number of bookings, with a consistent upward trend from the beginning of the year until mid-year, followed by a slight decline towards the year-end.

Weather Influence: Unsurprisingly, clear weather conditions correlate with higher booking rates, reflecting users' preference for biking in favourable weather.

Weekday Distribution: Thursdays through Sundays see higher booking numbers compared to the early days of the week, indicating increased biking activity over weekends.

Holiday Impact: Bookings tend to decrease on holidays, which aligns with expectations as people may opt for leisure activities closer to home during these times.

Working Day vs. Non-Working Day: Interestingly, there's a relatively balanced distribution of bookings between working days and non-working days, suggesting that biking remains popular regardless of the day of the week.

Yearly Progression: The year 2019 experiences a significant uptick in bookings compared to the previous year, signalling positive growth and increasing popularity of the bike-sharing service.

These observations provide valuable insights into the factors influencing bike demand and can inform strategic decisions to optimize service offerings and cater more effectively to user preferences.

2. **Why is it important to use drop_first=True during dummy variable creation?**

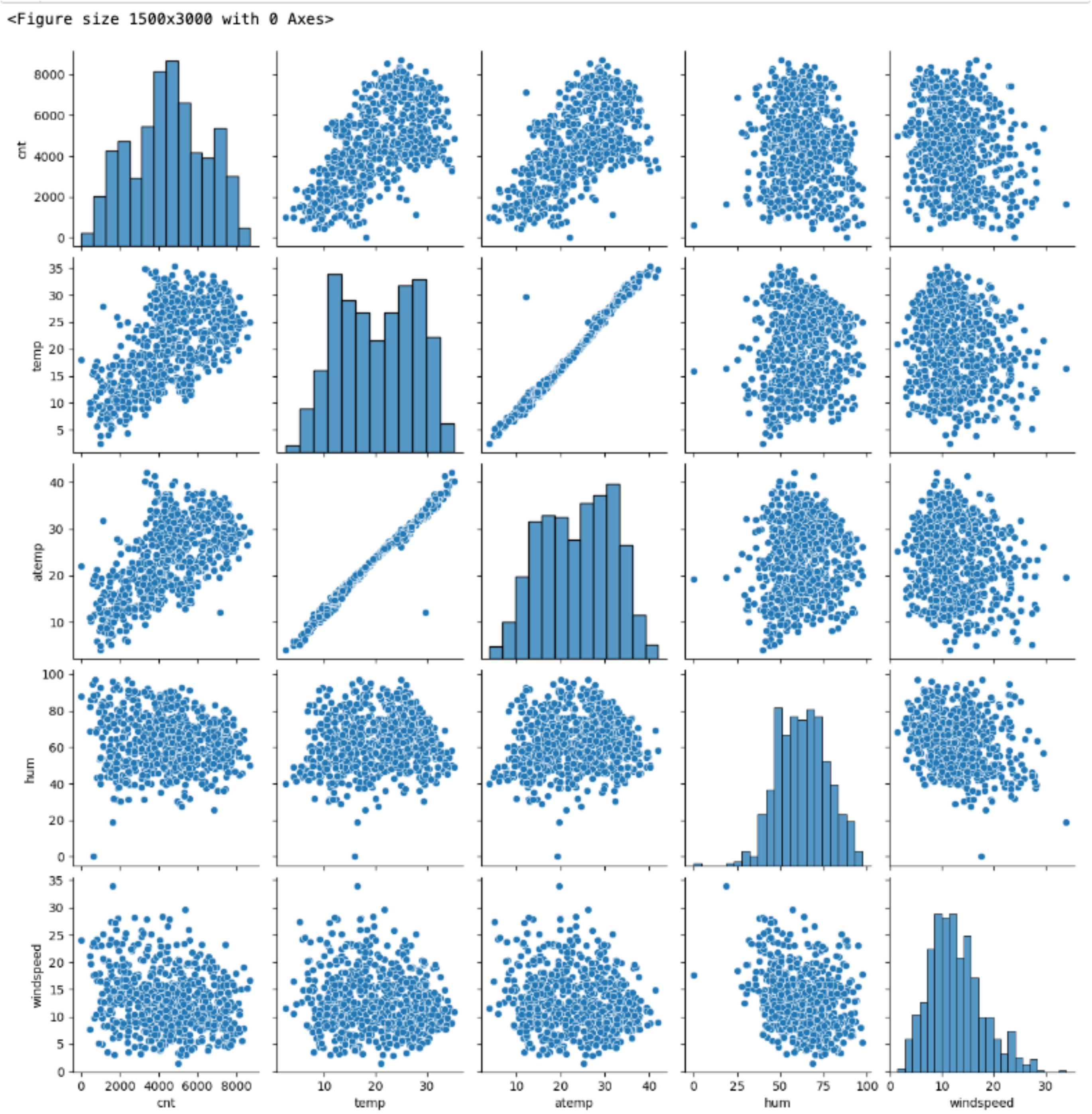
Utilizing drop_first=True during dummy variable creation serves two critical purposes:

Mitigating Multicollinearity: Including dummy variables for all categories of a categorical feature except one introduces multicollinearity into the model. This occurs because the information from the dropped category is inherently captured by the remaining categories. By dropping one category, we effectively avoid multicollinearity issues, which can distort model interpretation and affect its stability.

Enhancing Interpretability: Dropping the first category, often referred to as the reference category, enhances the interpretability of the model coefficients. With one category omitted, the coefficients of the remaining dummy variables represent the change in the dependent variable concerning the reference category. This simplifies the interpretation of the model's impact on the target variable and facilitates clearer insights into the relationships between the predictor variables and the outcome.

By incorporating drop_first=True, we not only enhance the robustness of the model but also streamline the interpretation process, thereby fostering more reliable and actionable insights from the analysis.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**



```
<Figure size 1500x3000 with 0 Axes>
```

The variables 'temp' and 'atemp' exhibit the strongest correlation with the target variable 'cnt' in comparison to the other variables.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

After building the model on the training set, Linear Regression assumptions were validated through Residual Analysis. The histogram showed a normal distribution of error terms around 0, indicating proper handling of Error Normal Distribution. Additionally, independence between error terms and predicted values was observed, while consistent variance along the fitted line suggested homoscedasticity.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

The three most influential factors in explaining the demand for shared bikes are temperature, year, and season.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a fundamental technique in supervised machine learning used for predicting continuous target variables. It involves fitting a linear relationship between the independent variables (predictors) and the dependent variable (target). There are two main types:

1. Simple Linear Regression:

In simple linear regression, there is only one predictor variable. The relationship between the predictor variable (x) and the target variable (y) is represented by the equation:

[y = b_0 + b_1x]

Here, (b_0) is the intercept and (b_1) is the coefficient or slope of (x).

2. Multiple Linear Regression:

When there are multiple predictor variables, the equation becomes:

[y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n]

In this case, (b_0) represents the intercept, while (b_1, b_2, ..., b_n) are the coefficients or slopes of the predictor variables (x_1, x_2, ..., x_n).

The goal of linear regression is to find the best-fitting line (or plane in multiple linear regression) that minimizes the sum of squared errors between the actual target values and the predicted values. This is achieved through various optimization techniques.

In summary, linear regression aims to model the relationship between the predictors and the target variable by fitting a linear equation, making it a powerful tool for prediction and inference in many real-world scenarios.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's Quartet comprises four distinct datasets that share almost identical summary statistics, yet exhibit significant differences when plotted and analysed visually. These datasets serve as a striking example of the limitations of relying solely on numerical summaries and the potential pitfalls of blindly applying regression models.

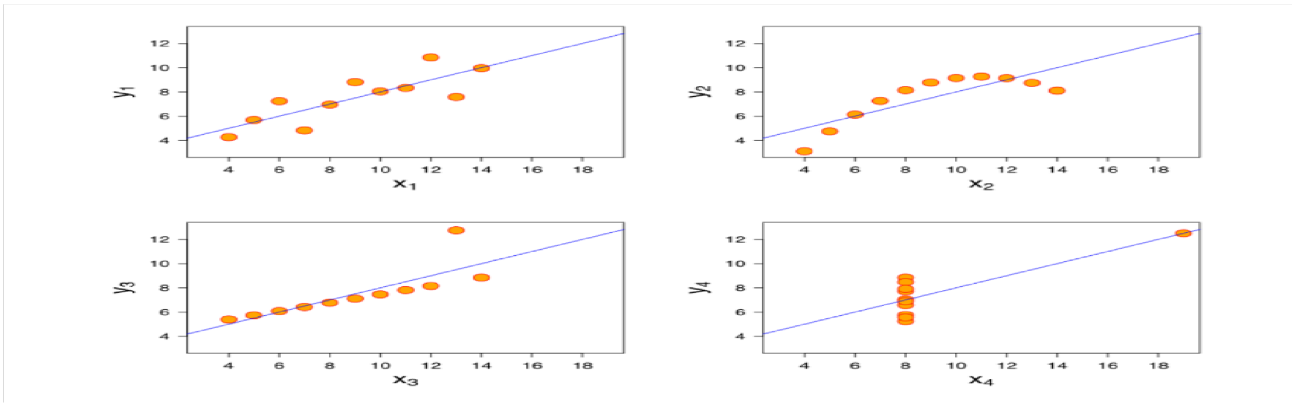
Each dataset within Anscombe's Quartet displays unique characteristics:

1. The first dataset demonstrates a clear linear relationship between the independent variable (X) and the dependent variable (y), making it suitable for linear regression analysis.

2. Conversely, the second dataset lacks a linear relationship between (X) and (y), rendering linear regression inappropriate for modelling.

3. The third dataset contains outliers, which can distort the results of linear regression models and undermine their accuracy.

4. The fourth dataset features a high leverage point, leading to a disproportionately influential effect on the correlation coefficient.



In essence, Anscombe's Quartet underscores the importance of visualizing data through plots and graphs before applying regression algorithms. By highlighting the limitations of numerical summaries alone, it emphasizes the critical role of exploratory data analysis in understanding the underlying patterns and relationships within datasets, ultimately guiding more informed model-building decisions.

3. **What is Pearson's R?**

Pearson's (r), also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1, where:

- (r = +1) indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.

- (r = -1) indicates a perfect negative linear relationship, implying that as one variable increases, the other variable decreases proportionally.

- (r = 0) indicates no linear relationship between the variables.

Pearson's (r) is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is sensitive to outliers and assumes that the relationship between variables is linear and that the data is normally distributed.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling refers to the process of transforming data so that it fits within a specific scale. It's a crucial preprocessing step aimed at ensuring that the features of the data are comparable and do not disproportionately influence the learning algorithm due to differences in their magnitudes, units, or ranges. Without scaling, algorithms may prioritize features with larger magnitudes, leading to inaccurate modelling.

There are two common types of scaling methods: normalized scaling and standardized scaling.

1. **Normalized Scaling:** This method scales the features to a range between 0 and 1 (or -1 to 1) based on their minimum and maximum values. It's particularly useful when the features have varying scales. Normalized scaling is sensitive to outliers and is suitable when the distribution of data is unknown.

2. **Standardized Scaling:** Unlike normalized scaling, standardized scaling uses the mean and standard deviation of the features to transform them. It ensures that the scaled features have a mean of 0 and a standard deviation of 1, resulting in a standardized distribution. Standardized scaling is robust to outliers and is preferred when the data follows a normal distribution.

Key differences between normalized scaling and standardized scaling include:

- Normalized scaling uses the minimum and maximum values, while standardized scaling uses the mean and standard deviation.

- Normalized scaling scales values to a specific range (0 to 1 or -1 to 1), whereas standardized scaling does not impose any bounds.

- Normalized scaling is sensitive to outliers, whereas standardized scaling is not significantly affected by outliers.

- Normalized scaling is suitable for unknown distributions, while standardized scaling is ideal for normally distributed data.

- Normalized scaling is also known as scaling normalization, while standardized scaling is referred to as Z Score Normalization.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

Sometimes, the value of VIF (Variance Inflation Factor) can become infinite, which typically occurs due to perfect multicollinearity among independent variables. VIF measures the extent to which the variance of an estimated regression coefficient is increased because of collinearity. When two or more independent variables are perfectly correlated, the VIF becomes infinite. The formulation of VIF is based on (frac{1}{1 - R^2}), where (R^2) is the coefficient of determination between the independent variable of interest and the remaining independent variables. In cases of perfect correlation, (R^2) becomes 1, resulting in a division by zero, leading to an infinite VIF value.

To address this issue, it's necessary to eliminate one of the variables causing perfect multicollinearity from the dataset. By removing redundant variables, we can mitigate multicollinearity and ensure that the VIF values are within an acceptable range. Typically, a VIF greater than 10 indicates high multicollinearity, while values above 5 should also be carefully examined and addressed.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

A Quantile-Quantile (Q-Q) plot is a graphical method used to compare two probability distributions by plotting their quantiles against each other. It helps assess whether a set of data likely originated from a specific theoretical distribution, such as a normal, exponential, or uniform distribution.

In linear regression, Q-Q plots are valuable tools for several reasons. Firstly, they aid in determining whether two datasets have similar distributions. When datasets are comparable, the Q-Q plot tends to exhibit greater linearity. This linearity assumption can be further validated with scatter plots. Additionally, linear regression analysis requires all variables to be multivariate normal, and Q-Q plots can effectively confirm this assumption.

The importance of Q-Q plots in linear regression lies in their ability to:

- Confirm whether both the training and test datasets are drawn from populations with the same distribution. Detect shifts in location, scale, symmetry changes, and the presence of outliers. Compare the distributional aspects, such as shifts in location and scale, between two datasets.

- Assess whether both datasets share a common distribution, location, scale, distribution shape, and tail behaviour.

Q-Q plots offer the advantage of being applicable to datasets of varying sizes and provide insights into the similarities and differences between distributions, aiding in robust model validation and interpretation.