Assignment-based Subjective Questions

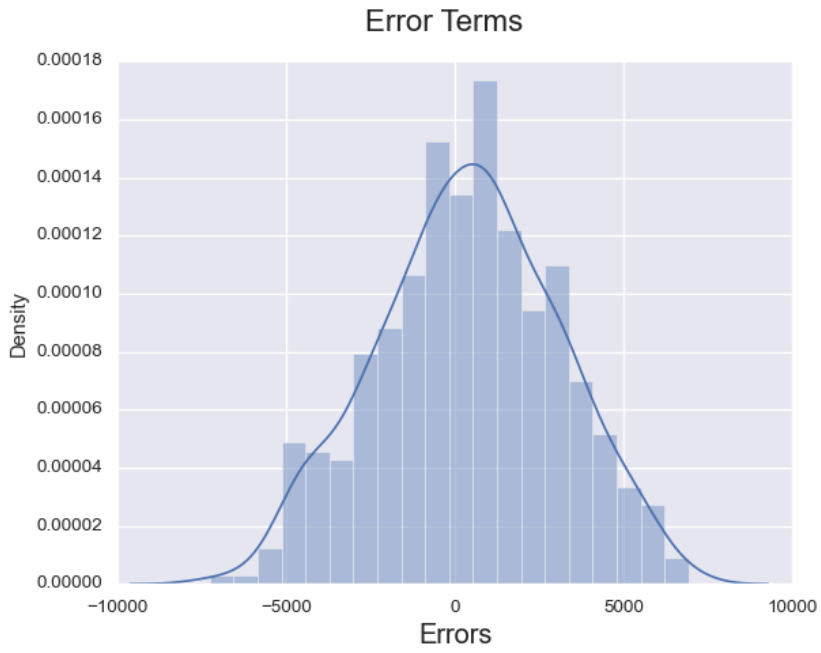
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)
There are many categorical variables like season, weathersit, holiday, working day, weekday, year and month.

- On **holidays**, there is less demand for bikes.
- Based on the **weather situation** also, there is varying demand. On a Clear, Few clouds, Partly cloudy, Partly cloudy (value =1), then the demand is the most while when there is heavy rain(value =4), there is no demand at all.
- As a **season**, fall is the best time of the year. Spring is when the demand goes down considerably.
- On **working days**, there is more demand for the bikes compared to holidays or weekends.
- The demand for bikes has almost doubled from 2018 to 2019 looking at **yr**.
- There is well over 300K in total demand from the **months** May to October.
- There is almost consistent demand for bikes all throughout the week looking at the **weekday** column**.**

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)
We use the pandas get_dummies method to convert a categorical variable to a numeric one. So if the categorical variable has 'k' level, then (k-1) levels are able to depict the info correctly. Hence the kth column is dropped using drop_first=True. This is to reduce the impact of collinearity between the independent variables.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)
The highest correlation with the target variable(cnt) is for the registered users at **95**%, then comes casual users at **67**%. For Temp/atemp, the % stands at **63**.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)
The residual analysis is used to test the assumptions of the linear model.
Basic assumption is that input and output variables have some sort of linear relationship. But the assumptions to get the output for all populations correct is related to the error/residual.
1. Error terms are normally distributed with mean zero
2. The error terms should not be dependent on one another
3. Error terms have constant variance (homoscedasticity)



The above chart obtained as part of the residual/error analysis shows that our assumptions are correct as its a normally distributed error with mean at 0.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Looking at coefficient values of the final model, we can make a call on which are the top features since all the features were scaled. Therefore **casual**(coeff=1063.4212) , **year** (coeff = 792.8863) and **workingday**(coeff = 692.3536) comes out as top 3 features. With **atemp**(coeff = 426.8082) as well as Winter **season**(coeff = 543.45) also having significant impact.

General Subjective Questions

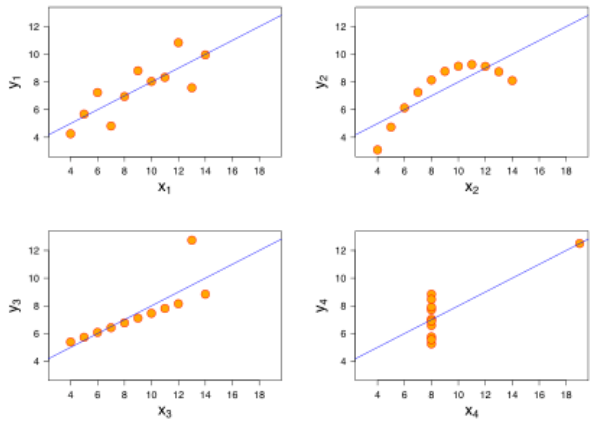
Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>
Linear regression is a supervised machine learning algorithm in which the relationship between a numerical dependent variable is analyzed using independent or predictor variables. Example: The rent(dependent variable) of a home based on area, location, amenities (predictor variables). Here we use historical data in order to study the pattern and a model is fed with this information so that we can come up with a formula to calculate the rent in a future case. Hence we are trying to get the coefficients of the formula where y=c + m1x1 + m2x2 + … + mnxn. Here x1, x2…xn are the predictor variables and y is the output variable. The coefficients are m1,m2,...mn. If we have only one predictor variable, then it's called a Single Linear regression model. There are many predictor variables, then it's called a Multiple Linear Regression. The assumption made while building a linear regression model is that the variables have a linear relationship with the output variable. When

building a multiple Regression model, there are considerations made to get the right mix of variables so that we do not overfit/underfit the model. For doing this we need to ensure the variable selection or feature selection is made keeping in mind the collinearity between them, essentially if 2 variables are having the same impact on output, then we keep only one of them.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>
Anscombe's quartet is a scatter plot representation of 4 dataset which has the identical summary statistics in terms of mean, variance, correlation coefficient yet their graphs show a different pattern for each. It reveals the importance of doing exploratory data analysis and visualizing the data. This helps to understand trends, patterns and outliers in the data. The datasets were created by the statistician Francis Anscombe in 1973. Each dataset has 11 x-y pairs of data.



Data-set (x1,y1) — consists of a set of (x,y) points that represent a linear relationship with some variance.
Data-set (x2,y2) — shows a curve shape but doesn't show a linear relationship
Data-set (x3,y3) — looks like a tight linear relationship between x and y, except for one large outlier.
Data-set (x4,y4) — looks like the value of x remains constant, except for one outlier as well.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>
The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. It is known by many names like Pearson's r, Bivariate correlation, Pearson product-moment correlation coefficient (PPMCC) and The correlation coefficient.
Depending on the value of r, the degree of correlation is interpreted as
Perfect: Values near ±1 indicate a perfect correlation, where one variable's increase (or decrease) is mirrored by the other.

High Degree: Values between ±0.50 and ±1 suggest a strong correlation.
Moderate Degree: Values between ±0.30 and ±0.49 indicate a moderate correlation.
Low Degree: Values below +0.29 are considered a weak correlation.
No Correlation: A value of zero implies no relationship.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>
Feature scaling is a preprocessing step in machine learning models where all the independent variables are brought in the same numeric range or normalized. This ensures that the magnitude of the different variables are considered appropriately by the model. The performance of the algorithm also improves when scaling is carried out. Numerical instability can be prevented by avoiding significant scale disparities between features. Scaling features means ensuring that each characteristic is given the same consideration during the learning process. Without scaling, bigger scale features could dominate the learning, producing skewed outcomes. This bias is removed through scaling.
Normalized scaling : This uses the min and max value of the dataset. The data is brought within the range of 0-1.
Standardized scaling: This uses the Z-score. It brings all the data with a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>
Variance Inflation Factor (VIF) is a measure of the amount of multicollinearity in regression. Two variables are said to be perfectly collinear when increase in one causes a similar increase or decrease in the other. Multicollinearity comes into picture since we have multiple variables in a regression model. So the degree of collinearity between the multiple variables can be explained by the VIF value.
The formula for VIF is :

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

$R_i^2 = \text{Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones}$

So when VIF = 1, the variables are not correlated.
but when VIF = ∞, then R-square is 1, which means that there is some independent variable which is perfectly correlated to another independent variable. This must be eliminated from the dataset in order to have proper results.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>
The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions. Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same. It is also used in the post-deployment scenarios to identify covariate shift/dataset shift/concept shift visually.
