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? (3 marks)

[Answer]

As per analysis of the categorical variables from the dataset using box plot

- The season box indicates that more bikes rented during fall season
- The year 2019 indicates that more bikes rented
- The working day & holiday box indicate that more bikes rented during normal working days than on weekends or holidays.
- The weathersit box plots indicates that more bikes are rent during Clear Sky days

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

[**Answer**]

If there is a column which has 10 unique categorical values or labels, using pd.getdummies() we convert them into a binary vector which makes 10 columns.

if drop_first is true it removes the first column which is created for the first unique value of a column.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

[**Answer**]

By looking at the pair plot temp variable has the highest (0.84) correlation with target variable 'cnt'.

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

(3 marks)

[**Answer**]

we must first make sure that four assumptions are met:

Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y. If you try to fit a linear relationship in a non-linear data set, the proposed algorithm won't capture the trend as a linear graph, resulting in an inefficient model.

if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data. The residuals (error terms) are independent of each other. It has no correlation between the consecutive error terms of the time series data.

Durbin-Watson (DW) statistic test. The values should fall between 0-4. If DW=2, no auto-correlation; if DW lies between 0 and 2, it means that there exists a positive correlation.

No Multicollinearity - (Variance Inflation Factor). VIF<=4 implies no multicollinearity, whereas VIF>=10 implies serious multicollinearity.

Homoscedasticity: The residuals have constant variance at every level of x. Create a scatter plot that shows residual vs fitted value. If the data points are spread across equally without a prominent pattern, it means the residuals have constant variance (homoscedasticity)

Normality: The residuals of the model are normally distributed. The last assumption that needs to be checked for linear regression is the error terms' normal distribution.

Check the assumption using a Q-Q (Quantile-Quantile) plot. If the data points on the graph form a straight diagonal line, the assumption is met.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

[**Answer**]

Significant variables to predict the demand for shared bikes:

- temp
- yr_2019
- Season(Spring,Summer,Winter)
- months(September, July)
- weathersit(Light Snow, Mist Cloudy)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

[Answer]

Linear Regression is machine learning algorithm based on supervised learning. It uses regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), which is called linear regression.

Linear regression uses a traditional formula y = mx + b or y = a0 + a1x
y – Dependent variable
x – Independent variable
m – Slope of line
b – Intercept of the line

The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients. The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals.

Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line (a0, a1 => xi, b) by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.

2. Explain the Anscombe's quartet in detail. (3 marks)

[Answer]

Anscombe proved the importance of the graphs with some sample datasets, though numerical calculations looks good it's completely wrong when we compare with graphical representations.

The basic thing to analyze about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, it appears very different when graphed. Each dataset consists of eleven (x,y) points.

Francis Anscombe demonstrated the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson's R? (3 marks)

[**Answer**]
In statistics, the Pearson correlation coefficient referred as Pearson's R(PPMCC).

It is the covariance of two variables, divided by the product of their standard deviations

The Pearson's correlation coefficient varies between -1 and +1

Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

Python API for pearson's r
from scipy import stats
>>> stats.pearsonr([1, 2, 3, 4, 5], [10, 9, 2.5, 6, 4])
(-0.7426106572325057, 0.1505558088534455)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   (3 marks)

[**Answer**]
The scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

In most of the times if scaling is not done, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units so it will lead to incorrect modelling.

Normalization typically means rescales the values into a range of [0,1]

Normalisation	Standardisation
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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

(3 marks)

[**Answer**]

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

(3 marks)

[**Answer**]

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.