**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)

- Almost 68.6% of the bike booking was happening during Clear weather with a median of close to 5000 bookings (for two years).
- This was followed by Misty with 30% of the total booking.
- It indicates that the weathers does show some trend towards the bike bookings, and it can be a good predictor for the dependent variable.
- The current data frame does not have any data where the weather is Heavy RainSnow
- Almost 32% of the bike booking were happening in Fall with a median of over 5000 bookings (for two years).
- It is followed by Summer & Winter with 27% & 25% of total booking.
- It indicates that the season can be a good predictor of the dependent variable.
- Almost 10% of the bike booking was happening in the months' May to Sep with a median of over 4000 bookings per month.
- It indicates that the month has some trend for bookings and can be a good predictor for the dependent variable.
- Almost 97% of bike rentals are happening during non-holiday time.
- Weekday variable shows the very close trend (between 13.5%-14.8% of total booking on all days of the week)
- Weekday Medians is between 4000 to 5000 bookings. This variable can have some or no influence on the predictor.
- Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 bookings (for two years).
- It indicates that the workingday can be a good predictor of the dependent variable

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

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Temp(temperature) has the highest correlation with the target variable.

**4. How did you validate the assumptions of Linear Regression after building the model on the training set?** (3 marks)

<u>Error terms must be normally distributed</u>

It is checked by plotting histogram of the error terms. The error terms must be normally distributed.

It can also be checked using a Q-Q (Quantile-Quantile) plot. If the data points on the graph form a straight diagonal line, the assumption is met.

<u>Homoscedasticity</u>

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). Otherwise, if a funnel-shaped pattern is seen, it means the residuals are not distributed equally and depicts a non-constant variance (heteroscedasticity).

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

Independent variable having the largest absolute value for its standardized coefficient.

1) Temperature (0.4923)
2) LightSnow (-0.2856)
3) Year (0.2338)

*Final Equation:*

cnt = 0.2036 + 0.2338yr + 0.4923temp - 0.1498windspeed - 0.0680spring + 0.0467summer + 0.0831winter - 0.0486Jul + 0.0721Sep - 0.2856LightSnow - 0.0816MistCloudy-0.0451Sun

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

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

Simple Linear Regression equation is:

$y = mx + c$

m is slope of line

c is y-intercept of line

x is independent variable.

y is dependent variable.

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

| Sno. | X1, Y1 | X2, Y2 | X3, Y3 | X4, Y4 |
|---|---|---|---|---|
| 1 | 10.0, 8.04 | 10.0, 9.14 | 10.0, 7.46 | 8.0, 6.58 |
| 2 | 8.0, 6.95 | 8.0, 8.14 | 8.0, 6.77 | 8.0, 5.76 |
| 3 | 13.0, 7.58 | 13.0, 8.74 | 13.0, 12.74 | 8.0, 7.71 |
| 4 | 9.0, 8.81 | 9.0, 8.77 | 9.0, 7.11 | 8.0, 8.84 |
| 5 | 11.0, 8.33 | 11.0, 9.26 | 11.0, 7.81 | 8.0, 8.47 |
| 6 | 14.0, 9.96 | 14.0, 8.10 | 14.0, 8.84 | 8.0, 7.04 |
| 7 | 6.0, 7.24 | 6.0, 6.13 | 6.0, 6.08 | 8.0, 5.25 |
| 8 | 4.0, 4.26 | 4.0 3.10 | 4.0, 5.39 | 19.0, 12.50 |
| 9 | 12.0, 10.84 | 12.0, 9.13 | 12.0, 8.15 | 8.0, 5.56 |

| | | | | |
|---|---|---|---|---|
| 10 | 7.0, 4.82 | 7.0, 7.26 | 7.0, 6.42 | 8.0, 7.91 |
| 11 | 5.0, 5.68 | 5.0, 4.74 | 5.0, 5.73 | 8.0, 6.89 |

All the summary statistics are identical:
The average x value is 9 for each dataset
The average y value is 7.50 for each dataset
The variance for x is 11 and the variance for y is 4.12
The correlation between x and y is 0.816 for each dataset
A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3.
But when these values are graphed (scatter plot), the relationship is totally different.

**3. What is Pearson's R?** (3 marks)
The Pearson correlation measures the strength of the linear relationship between two variables.

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

It is the ratio between the covariance of two variables and the product of their standard deviations.

r(xy) = s(xy) / s(x)*s(y)
r(xy) is sample correlation co-efficient
s(xy) is the sample covariance.
s(x) and s(y) are the sample standard deviations.

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?** (3 marks)
Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.
Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Normalization
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
X' = (X-X(min))/ (X(max) – X(min))

Standardization (Z-score Normalization)
The general method of calculation is to determine the distribution mean and standard deviation for each feature. Next, we subtract the mean from each feature and then divide it by standard deviation.

Z = (X-mu)/sigma
Where Mu is mean and sigma is standard deviation

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?** (3 marks)
VIF = 1/(1-R-squared)
If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.
A large value of VIF indicates that there is a correlation between the variables. If there is perfect correlation, then VIF = infinity.

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.** (3 marks)
QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values.It determines how many values in a distribution are above or below a certain limit.
If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line.

**References: Upgrad Study Material, Wikipedia.**