Darshan Patel
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)

➔ 1. By the categorical variables, we can infer that the count distribution in the year 2019 is better than the year 2018 with the highest count of 8714.

2. Further, we can infer that bikes are least rented in spring and increase in summer. After summer, there is a sudden increase in bike renting during the fall with a slight decline in rents during winter.

3. Ridership is based on weather as well. When the weather is clear, the number of ridership is high. There is a slight decline in rents during mist + cloudy weather following with light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds weather. And there is not a single rent placed during heavy Rain + Ice Pallets + thunderstorm + mist, snow + fog weather.

4. More we can say that the bikes are rented more during working days as compared to holidays.

5. It can be seen that bike rents are low during January with an increasing graph till June. Then the rent cont average is around 5000 from June to October, and then again there is a decrease in counts till the year-end.

6. On daily basis, we can observe that the average is around 4200 bike rent counts. We can also visualize some of these categorical features parallelly by using the hue argument.

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

➔ Whether to get k-1 dummies out of k categorical levels by removing the first level.

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

➔ Temperature is the only numerical variable that has 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)

➔ We will be validating the assumptions of liner regression after building the model on the training set base on Normal distribution of error terms and .Little or No autocorrelation in the residuals

1. Normal distribution of errors terms: The error (residuals) follow a normal distribution. However, a less widely known fact is that, as sample sizes increase, the normality

assumption for the residuals is not needed. More precisely, if we consider repeated sampling from our population, for large sample sizes, the distribution (across repeated samples) of the ordinary least squares estimates of the regression coefficients follow a normal distribution. The histogram plot in the “Error (residuals) vs Predicted values” also shows that the errors are normally distributed with mean close to 0.

2. Little or No autocorrelation in the residuals: Autocorrelation can be tested with the help of Durbin-Watson test. From the summary of our model 12 note that the value of Durbin-Watson test is 2.01 when the value of Durbin-Watson is equal to 2, r takes the value 0 from the equation 2*(1-r),which in turn tells us that the residuals are not correlated.

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

➔ Temperature , year and spring are the top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

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

➔ 1. Let us start by loading the necessary libraries. Firstly, we are importing Pandas which is the most popular python library for data exploration, manipulation and analysis. Please follow this document to install Pandas. We are importing Mathplotlib for multiplatform data visualization.

➔ 2. Next steps we are going to load the dataset, read the data into a data frame and display the head (top 5 rows). Also, we can see the total number of rows

➔ 3. Preparing X and y

➔ 4. Splitting data into train and test

➔ 5. Performing Linear Regression

➔ 6. Coefficients calculation

➔ 7. Making predictions

➔ 8. Model evaluation (Plot Actual vs Predicted)

➔ 9. Model evaluation (Plot Error terms)

➔ 10. Checking mean square error and R square

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. They to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

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

➔ Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

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

➔ Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

1. 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.
2. Standardization is another scaling technique where the values are centered on the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

➔ An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).In other words 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.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

➔ Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

1. Use of Q-Q plot: a) it can be used with sample sizes also. b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
   It is used to check following scenarios:
   If two data sets —

i. come from populations with a common distribution
ii. Have common location and scale
iii. Have similar distributional shapes
iv. Have similar tail behavior