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? (3 marks)

Answer: Following points we could infer from the data set:

- The bike demand is high when weather is clear and few clouds. However, the demand is less in case of Light snow and light rainfall. No conclusion can be made to Light Snow, Light Rain due to lack of data.
- Bike demand in the season of 'fall' is high when compared with other seasons and is consistent for 2018 and 2019 years.
- Bike demand is significantly high when it is not a holiday against when it is a holiday.
- There is increase in bike demand during working day against non-working day for the year 2019 when compared to year 2018.
- Bike demand during May to July months is high for the year 2018. Also, the demand for Bike during June to September months is high for the year 2019.
- The demand of bike is almost same throughout the weekdays for the year 2018. However, for the year 2019, the demand for bike is slightly increasing during weekday 3, 4 and 5.

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

Answer: “drop_first=True” is important to use, as it helps in reducing the extra column created during dummy variable creation. This helps to keep the model optimal and hence reduces the correlations created among dummy variables. This is very helpful when we are dealing with big datasets.

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

Answer: As per pair plot and heat map highest correlation with target variable ‘cnt’ is with ‘temp’ and ‘atemp’ i.e. 0.65. The variable ‘yr’ with 0.59 followed by ‘season_3’ with 0.37.

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

Answer: For validation of linear regression model on train set we can check following:

- Check the value of R2 as it should vary between 0 to 1. If R2 value is close to +1. It means nearly all data points are explained by linear regression model and vice versa. Any mode with R2>0.7 is considered to be a good.

- Check the value of adj.R2. It shall always be positive and less than R2. Adding relevant predictors or target variable tends to increase adj.R2 and vice- versa. This is where adj. R2 comes in picture to explain data points by model.
- The variables p values should be below 0.005 i.e., 5% for coefficients to be significant.
- Check for residual analysis. The distributions of residuals should be normally distributed to rely on inferences if any.

Question 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: Based on final model top three features contributing significantly towards explaining the demand are:

- Actual Temperature (0.5755)
- weathersit_3 : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.2766)
- year (0.2335)

General Subjective Questions

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

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variable or variables. It is mostly used for finding relationship between variable and forecasting. The whole idea is to get a best fit in line against all data points usually with RSS or OLS method. Linear regression best fit line is represented by equation:

$y = c+ m1X1+m2X2+ ..... + mnXn+e$, here y is value of the “target variable”, c is “intercept” and e is the “error”.

m1 is coefficient/slope if variable x1 and so on up to n

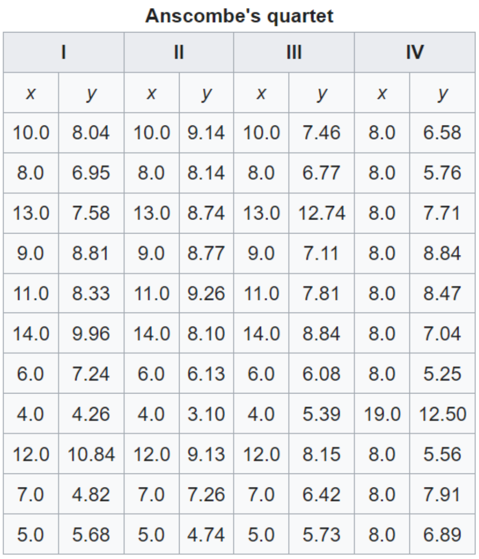
Following are assumptions in linear regression.

- It is assumed that target and dependent variables are in linear relationship. Assumptions about residuals:
- Residuals are in Normal distribution with zero mean value.
- Residual terms have same variance. i.e. homogeneity in error terms.
- Residual terms are independent to each other.
- Independent variables linearly independent to each other. There is no multicollinearity in data.

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

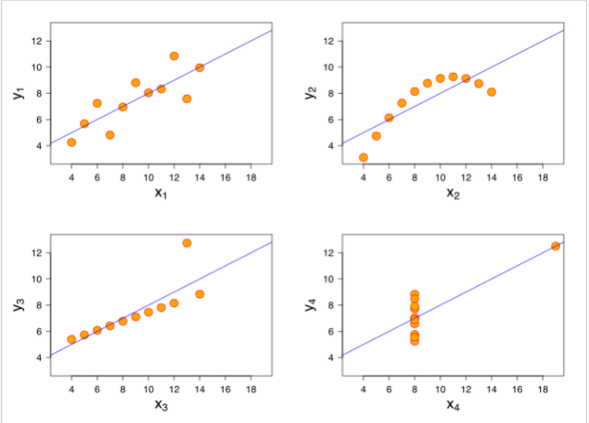
Answer: Anscombe’s quartet can be defined as group of four data sets which are nearly identical in simple descriptive statistical details, but there are peculiarities in data set that fool the regression model if build. They have very different distributions and appear differently when plotted in scatter plots. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it. The effect of outliers and other influential observations on statistical properties. He proved power of data visualization.

Example taken from the Wikipedia for better expandability:



Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

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

Answer: In statistics, the Pearson's correlation coefficient (R) is the product-moment correlation coefficient, the bivariate correlation is a measure of linear correlation between two data set. It is calculated by following formula as mentioned below:

- The value of R ranges from -1 to +1.
- The sign of R indicates direction.
- Positive sign means direct relation between two datasets and vice versa.
- The value of R indicates strength of correlation.
- Value close to one show strong correlation and close to zero is weaker one.

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

Answer: Feature scaling is method to normalize the range of independent variable or features of dataset within range. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. It also helps in speeding up calculation for algorithm. Most of the times, 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 hence incorrect modelling. To deal with this issue, scaling is done to bring all variables at same level of magnitude. It is important to note that scaling just affects the coefficients and not any other parameter as t-statistic, F- statistic, p- values and R-squared, etc.

Generally Scaling can be performed by two types i.e.,
Normalized Scaling: It is a technique in which values are shifted and rescale so that they end up ranging between zero and one. It is also called as Min-Max scaling.

- Formula for Normalized Scaling is X' =((X-Xmin)/(Xmax-Xmin))
- When X is min value, numerator shall be zero. Hence X' shall be zero.
- When X is max, then X' shall be 1.

- If X is in between min and maximum, value of X' shall be in between 0 and 1 respectively. It is used usually when outliers are present in the data points and do not follow Gaussian curve.
- It supports greater overall database organization. Data consistency with in the database

Standardized Scaling: It is another technique of feature scaling where values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and resultant distribution has a unit standard deviation.

- It is calculated by $X'=((X-\mu)/\sigma)$. Here, μ is mean of feature and σ is standard deviation of feature values.
- It is used when data points are distributed on Gaussian curve. It is useful when data points have negative values.

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

Answer: If there is a perfect correlation, then variance inflation factor shall be come infinite. In case of perfect correlation we shall have R2 as one. Which lead to VIF = (1/(1- R2)) as infinite. To solve this problem we need to drop one of the variable from dataset which is causing this perfect multicollinearity.

An infinite VIF indicates that the corresponding variable may be expressed exactly by a linear combination of other variable (which show an infinite VIF as well). As a referral method to identify multicollinearity between variables we calculate VIF.

IF VIF = 1, there is absolutely no correlation between two variables. Ideally VIF between 1 to 5 is considered good for variable to be used for model building. If VIF is between 5 to 10, variable should be dropped out while model building.

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

Answer: A Q-Q plot are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45° angle is plotted on the Q-Q plot: if the two data sets come from a common distribution, the points will fall on that reference line.
If the two distributions being compared are similar, the Q-Q plot will approximately lie on the line y=x. If the distribution is are linearly related, the points in the Q-Q plot will approximately lie on a line. But not necessarily on the line y=x. Q-Q Plots can also be used as a graphical means of estimating parameters in a location-scale family of distribution.

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