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?
    a. Here are some of the inferences I made from my analysis of categorical variables from dataset on the dependent variable(Count)-
        i. Fall has the highest median,which is expected as weather conditions are most optimal to ride bike followed by summer.
        ii. Median bike rents are increasing year on as year 2019 has a higher median then 2018, it might be due to the fact that bike rentals are getting popular and people are becoming more aware about environment.
        iii. Overall spread in the month plot is reflection of season plot as fall months have higher median.
        iv. People rent more on non-holidays,so reason might be they prefer to spend time with family and use personal vehicle instead of bike rentals.
        v. Overall median across all days is same but spread for Saturday and Wednesday is bigger may be evident that those who have plans for Saturday might not rent bikes as it a non-working day.
        vi. Working and non-working days have almost the same median although the spread is bigger for non- working days as people might have plans and do not want to rent bikes because of that.
        vii. Clear weather is the most optimal for bike renting,as temperature is optimal, humidity is less, and temperature is less.
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
    a. To avoid multi collinearity(If, we don't drop ,dummy variables will be correlated) and affects the model adversely.
    b. To avoid redundant features
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    a. Count(Target Variable) has significantly high correlation with Temperature(temp)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    a. Residual Errors follow normal distribution
    b. Maintains Linear Relation between dependent variable (Test and Predicted)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    a. Temperature(0.4354)
    b. Weather Situation-Light and Snowy(0.2837)
    c. Year(0.2461)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    a. 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 variables. It is mostly used for finding out the relationship between

variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

b. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.
c. An example is let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.
d. In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.
e. One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.
f. Linear regression is used to predict a quantitative response Y from the predictor variable X.
g. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
h. Y=theta1 + theta2*x
i. While training the model we are given :
   x: input training data (univariate – one input variable(parameter))
   y: labels to data (supervised learning)
    i. When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.
       θ1: intercept
       θ2: coefficient of x

2. Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

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

a. **Anscombe's Quartet** can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties**.There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.

b. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.

4. It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed. One of these, the Datasaurus Dozen, consists of points tracing out the outline of a dinosaur, plus twelve other data sets that have the same summary statistics. Datasaurus Dozen was created by Justin Matejka and George Fitzmaurice. The process is described in their paper "Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing".
5. The Datasaurus Dozen proves us as much as Anscombe Quartet why visualizing our data is important as summary statistics can be the same, while data distributions can be very different.

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

In statistics, the Pearson correlation coefficient — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data. It is the ratio

between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

For a population

Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson

correlation coefficient. Given a pair of random variables {\displaystyle (X,Y)} , the formula for ρ

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- is the covariance
- is the standard deviation of
- is the standard deviation of

The formula for rho can be expressed in terms of mean and expectation. Since

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

the formula for rho can also be written as

$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Practical issues

Under heavy noise conditions, extracting the correlation coefficient between two sets of stochastic variables is nontrivial, in particular where Canonical Correlation Analysis reports degraded correlation values due to the heavy noise contributions. A generalization of the approach is given elsewhere.

In case of missing data, Garren derived the maximum likelihood estimator.

Mathematical Properties:

The absolute values of both the sample and population Pearson correlation coefficients are on or between −1 and 1. Correlations equal to +1 or −1 correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation). The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X).

A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. That is, we may transform X to a + bX and transform Y to c + dY, where a, b, c, and d are constants with b, d > 0, without changing the correlation coefficient. (This holds for both the population and sample Pearson correlation coefficients.) Note that more general linear transformations do change the correlation: see § Decorrelation of n random variables for an application of this.

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

It 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.

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 solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

If there is perfect correlation, then VIF = infinity. 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 (which show an infinite VIF as well).

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

   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. 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 degree 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 points in the Q–Q plot will approximately lie on the line y = x. If the distributions 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 distributions.

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.