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?

Answer: -

There are four categorical variables found in dataset that **is weathersit, mnth, season, weekday.** For which we have created dummy variables and added to set for further analysis.

As per the analysis done on linear model I have made following inferences.

1. **weathersit** has four subcategories which I assumed as **light, heavy, mist and clear.** And the share of bikes is high during the mist weather (i.e., Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist)
2. from **weekday** Wednesday is the day of the week on which the bike share demand is high as compared to other days.
3. Feb is the month of the year which has high demands for bike sharing as compare to other months.

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

Answer: -

Creating dummy variables from categorical variable is an important step from data preparation in a linear regression process.

If we have a categorical variable **m** with level **n** then it is advisable to create **n-1** dummy variables.

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.

Example: -

Furnishing status is a categorical variable (m) with 3 levels(n) as “furnished”, “semiFurnished” and “unfurnished”. so, we can create dummy variables as

status = pd.get_dummies(housing['furnishingstatus']) **OR** ---- gives 3 columns

status = pd.get_dummies(housing['furnishingstatus'**], drop_first=True**) --- gives 2 columns by deleting extra one

the first gives us n no of dummy variables where the second line will give us n-1 no pf dummy variables for furnishing status.

With second approach we have a set of dummy variables with reduced multicollinearity.

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

Answer: -

The **registered** column is having high correlation with our target variable (cnt).

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

Answer: -

I have performed the Residual Analysis on the Model built on training set.

In residual analysis we look at the assumptions that need to be hold true for linear regression. Most common assumption is we look at the distribution of errors terms. The error terms should be normally distributed to mean zero.

So, steps I performed are -

Step 1: Getting the predicted values for all data points of target variable y from train set. To get this values (y_train_pred) I have used the predict method from linear regression model object.

Y_trian_pred = lr_model.predict(x_train_sm)

Step 2: calculate the residual points (Error Terms)

Res= y_train – y_train-pred

Step 3: plot distribution of these residual points and observe its distribution

sns.distplot(res)

Step 4: Observed the distribution of error terms as it need to be normally distributed towards mean zero.

---

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

Answer: - From Final model I got seven features (Registered, Casual, Holiday, Wednesday, Mist, Feb, Yr) who significantly contribute towards demand of bike sharing.

From these Features the top 3 features that highly contributes to demand of bike shares are as follows: -

1. Registered - from the model summary it has a positive coefficient also it is highly correlated with count. That is if number of registered users increased then the demand for bike share will also increase.
2. Casual - after the registered the casual user counts are the one whose increase in value will increase in demand of bike shares
3. Wednesday – its an observation the Wednesday is the day of week on which demand for bike is seems high.

General Subjective Questions

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

Answer :-

Simple Linear Regression Algorithm Steps : -

1. Reading and understanding the data
2. Visualizing Data (pair plot and heatmap will give better visualization)
   a. Select the variable which is highly correlated with target variable
3. Build The model
   a. Define x and y for model building
   b. Split the data into train and test
   c. Build the model
   d. Look at the summary for low p-values , high r-squared, and low f-statistics for best fit model
4. Perform Residual Analysis (Check for normal distribution of error terms)
5. Make predictions on test data set

Multiple linear regression Algorithm Steps :-

1. Reading and understanding the data
   a. Reading data set and its various parameters like no of columns , rows, no of non null values, understanding the numeric plus categorical variables
   b. Visualizing the data – plot graphs to look at visualizations and corelations between variables
2. Data Preparation
   a. Perform EDA to clean the data , handle the format of data properly.
   b. Create Dummy variables from categorical variable
   c. Append dummy variables and drop originals from the main dataset
   d. Split data into Train and Test (Ratio can be 70-30/80-20)
   e. Rescaling the feature
   f. Look at the correlation among all the variables so that we can select the variables for model building
3. Building Model :- There two techniques to build the model
   a. Select the model building Technique (forward selection Or backword elimination)
      i. Forward selection :-
         1. Select one variable at a time and start building model with it.
         2. Add another feature and rebuild the model.
         3. Repeat until we have enough variables that fits our model.
      ii. Backword Elimination
         1. Include all variables at first (can be done manually or by using Automated way like RFE
         2. Build the model and look at the summary
         3. Drop the variables which are insignificant one by one
4. Perform Residual Analysis (validating assumption) : -
   a. Calculate Error Terms (residual points)

    b. Plot distributions of Error Terms
    c. Check if the distribution is normal with mean 0.
    d. If distribution is normal we are good for model evaluation
5. Make predictions on test dataset
6. Model Evaluation
    a. Calculating r2 for test and comparing with model if similar then the model is good.
    b. Understand the spread for r2

---

2. Explain the Anscombe's quartet in detail.

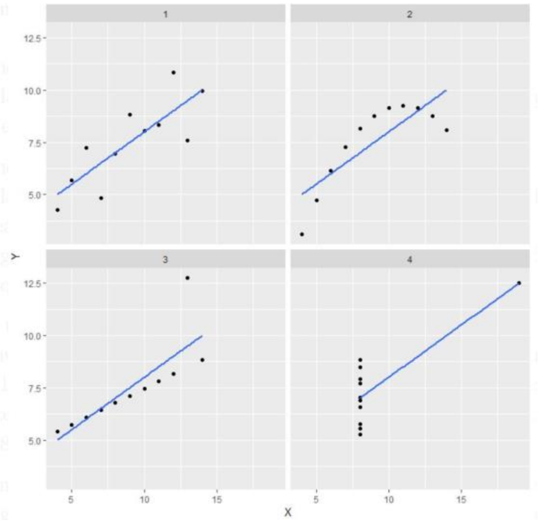
Answer: -

Anscombe's quartet **comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed**. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Usage Of Anscombe's Quartet: - It is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Example: -

Here is the output graph for Anscombe's quartet four dataset which are identical and consist of 11 points



Explanation of this output:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

---

3. What is Pearson’s R?

Answer: -

Correlation coefficients are used to measure how strong a relationship is between two variables.

There are several types of correlation coefficient, but the most popular is Pearson’s.

Pearson’s correlation (also called Pearson’s *R*) is a correlation coefficient commonly used in linear regression.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

If **Pearson’s R == 0** then there is no relation between two variables at all.

if **Pearson’s R == 1** then two variables share a strong positive correlation that means the value of y increases when x increases.

if **Pearson’s R == -1** then two variables share a strong negative correlation that means the value of y decreases when x increases.

Formula To Calculate Pearson’s R: -

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where,

r = correlation coefficient,

Xi= values of the x-variable in a sample,

Yi = values of the y-variable in a sample,

ȳ = mean of the values of the y-variable,

x̄ = mean of the values of the x-variable

For example: Up till a certain age, (in most cases) a child’s height will keep increasing as his/her age increases. i.e. (x as height and y as age) so the Pearson’s R should give correlation Coefficient value for x and y 1 or near to 1 (e.g., 0.99, 0.98, etc).

---

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: -

Scaling: - Scaling / 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.

There are two techniques by which we can scale the Feature/variables: -

1. Min-Max Scaling (normalized Scaling)

2. Standardized Scaling

Need of Scaling (why Scaling): -

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 must do scaling to bring all the variables to the same level of magnitude.

It also Speed ups the calculation in an algorithm.

difference between normalized scaling and standardized Scaling :-

- Difference by definitions and formulas -
   1. **Min-Max Scaling (normalized Scaling)**
      it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

      $$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

      Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

   2. **Standardized Scaling**
      Feature standardization makes the values of each feature in the data have zero mean and unit variance. Formula –

      $$x' = \frac{x - \bar{x}}{\sigma}$$

      Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.
- Normalization scaling scales values between [0,1] or [1,-1] where the standardization is values are not bounded to any range.
- Normalization scaling is much affected by outliers than of standardization one.
- Normalization scaling is chosen over standardized one when we don’t know about distributions.

Just to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example- cantered around 0 or in the range (0,1) depending on the scaling technique.

---

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

Answer: -

A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.

Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.

Formula for VIF calculation is –

VIF = 1/(1-R2)

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.

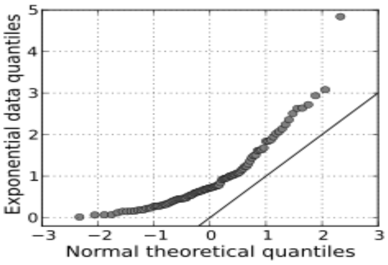
Answer :-

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.

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.

A Q Q plot showing the 45-degree reference line:



Importance: -

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

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.