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:

I have done analysis on categorical columns using the boxplot. Below are the few points we can infer from the visualization –

1. Fall season seems to have attracted more booking. Spring season having less demands as comparison to all the four season (Spring, Summer, fall and winter)
2. Booking count has increased drastically from 2018 to 2019
3. Maximum bookings have been done during the month of May, June, July, Aug, Sep and Oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
4. Clear weather attracted more booking which seems obvious.
5. Thu, Fri, have a greater number of bookings as compared to the start of the week.
6. Booking on holiday is little less than the other day, that may be because on holiday people want to spend time at home with family.
7. Booking seemed to be almost equal either on working day or non-working day.
8. 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of business.

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

Answer:
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.

Syntax :

drop_first: bool,

default False,

which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's say we have 3 types of values in one Categorical column, for eg. In our dataset season, it has three (1: spring, 2: summer, 3: fall, 4: winter)

Now we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So, we do not need 3rd variable to identify the C. In case of

season, if neither spring, summer and fall then it's obvious that it is winter. So, we can infer this using three values of season, so do not require to create fourth variable. Which reduce the overhead of the model created.

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

Answer:

'temp' and **'atemp'** variables shows highest correlation with the target variable.

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

Answer :

I validate the assumptions of Linear Regressions :

1. **Normality of error terms**

   Plot dist-plot with error term to check whether it is normally distributed of that. It is normally distributed

2. **Multicollinearity check**

   There should be insignificant multicollinearity among variables. Which we checked with VIF values of all the features that all are below 5. Which is desired. Similarly, adj. r-square value (.827) is also approximately same as r-square value (.830) of the final model. Which is good.

3. **Linear relationship validation**

   Linearity should be visible among variables, By plotting scatter plot, we found that there having linear relation with various features with target variable like season, temp, hum etc.

4. **Homoscedasticity**

   There should be no visible pattern in residual values. By plotting the residuals, we found there does not have any pattern in error term. Hence shows Homoscedasticity.

5. Independence of residuals
   No auto-correlation. When plotting the error terms, with selected features, I confirmed that there does not have any patterns or correlations

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:

Significant variables to predict the demand for shared bikes

- holiday
- temp
- windspeed
- Winter and Summer Season
- September
- Year (2019)
- weathersit(Light Snow, Mist + Cloudy)

Out of above holiday, windspeed, light snow and mist weather have negative impact on demand, which is intuitive.

General Subjective Question

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

Answer:
Linear regression is a statistical method used to model the relationship between a dependent variable (often called the response variable) and one or more independent variables (also known as predictors or features). The goal is to find the linear relationship that best predicts the dependent variable from the independent variables.
Mathematically the relationship can be represented with the help of following equation –

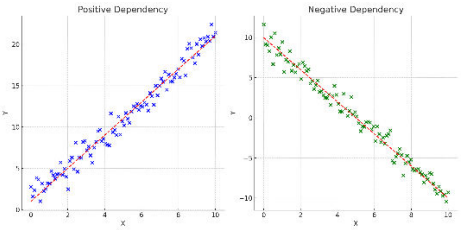
$$Y = mX + c$$

Here,
Y : is the dependent variable we are trying to predict.
X : is the independent variable we are using to make predictions.
m : is the slope of the regression line which represents the effect X has on Y
c : is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

The linear relationship can be positive or negative in nature as explained below−

a. Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases.

b. Negative Linear Relationship: A linear relationship will be called negative if independent increases and dependent variable decreases



Types of Linear Regression

1. Simple Linear Regression

Simple linear regression models the relationship between two variables by fitting a linear equation to observed data. The equation of the line is:

$$y=\beta0 + \beta1x + \epsilon$$

- y : dependent variable
- x : independent variable
- β0 : y-intercept (the value of y when x is 0)
- β1: slope of the line (the change in y for a one-unit change in x)
- ϵ: error term (the difference between the observed and predicted values)

2. Multiple Linear Regression

When there are multiple independent variables, the model is extended to:

$$y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\epsilon$$

Assumptions of Linear Regression

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: The residuals (errors) are independent.
3. **Homoscedasticity**: The residuals have constant variance at every level of xxx.
4. **Normality**: The residuals of the model are normally distributed.

Question 2. Explain the Anscombe's quartet in detail. (3 marks)
Answer:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but exhibit very different distributions and appear very different when graphed. It was constructed by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and to show how statistical properties can be misleading if not interpreted properly in the context of the data's distribution and visualization.

Anscombe's Quartet Dataset
The four datasets of Anscombe's quartet

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

The Four Datasets

Each of the four datasets in Anscombe's quartet consists of eleven (x,y)(x, y)(x,y) points. Despite having nearly identical statistical properties, the datasets are very different when graphed. Here are the common statistical properties for all four datasets:

- Mean of x
- Mean of y
- Variance of x
- Variance of y
- Correlation between x and y
- Linear regression line y=a+bx
- Coefficient of determination R-square of the regression

Detailed Examination

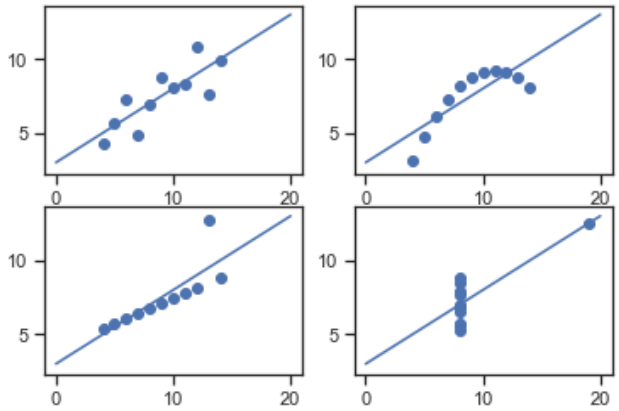
1. Summary Statistics

The summary statistics for the four datasets are:

- Mean of x for all datasets: 9
- Mean of y for all datasets: 7.5
- Variance of x for all datasets: 11
- Variance of y for all datasets: 4.125
- Correlation between xxx and y for all datasets: 0.816
- Linear regression line for all datasets: y=3+0.5x
- R square for all datasets: 0.67

Despite these similarities, the datasets are quite different when plotted.

1. *Graphical Representation*



1. **Dataset I**: The data points form a roughly linear pattern with some scatter around the regression line. This is a typical example of what might be expected for a dataset with these statistics.
2. **Dataset II**: The data points form a perfect curve, indicating a non-linear relationship. The linear regression line does not fit the data well, even though the summary statistics suggest a strong linear relationship.
3. **Dataset III**: The data points form a vertical cluster with one outlier. This outlier significantly influences the calculation of the regression line and the correlation coefficient, making the summary statistics misleading.
4. **Dataset IV**: The data points are mostly clustered with one outlier that is vertically aligned. This outlier skews the summary statistics, leading to a misleading representation of the data's distribution and relationship.

Importance of Anscombe's Quartet

Anscombe's quartet emphasizes several key points in data analysis:

1. **Graphing Data**: Visualizing data is crucial. Graphs can reveal patterns, relationships, and anomalies that are not apparent from summary statistics alone.
2. **Exploratory Data Analysis (EDA)**: Before applying any statistical methods, it is essential to explore the data through visualization and simple descriptive analysis to understand its structure and distribution.
3. **Contextual Interpretation**: Summary statistics can be identical across different datasets but may not provide a complete picture. The context and the nature of the data should always be considered when interpreting statistical results.
4. **Influence of Outliers**: Outliers can significantly affect summary statistics, correlation coefficients, and regression lines. Identifying and understanding the impact of outliers is important in data analysis.

By illustrating these points, Anscombe's quartet remains a powerful teaching tool in statistics, highlighting the necessity of graphical analysis and contextual understanding in data analysis.

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

Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between the variables. The coefficient is named after Karl Pearson, who developed it.

Formula

The Pearson correlation coefficient is calculated using the following formula:

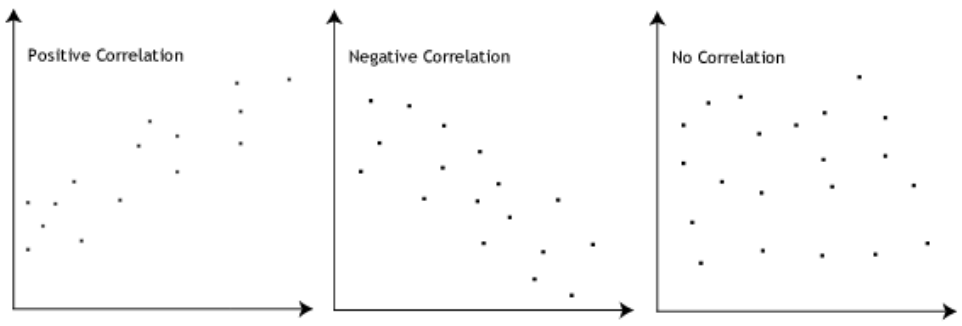
$$R = (\sum(xi-x)(yi-y))/ \text{Sqrt}(\sum(xi-x)(yi-y))$$

where:

- xi and yi are the individual sample points.
- x : is the mean of the x values.
- y : is the mean of the y values.

Interpretation

- r ranges from -1 to 1.
  - r=1: Perfect positive linear correlation.
  - r−1: Perfect negative linear correlation.
  - r=0: No linear correlation.
  - 0<r<1: Positive correlation.
  - −1<r<0: Negative correlation.



Key Points

1. **Direction**:
    - Positive r: As x increases, y tends to increase.
    - Negative r: As x increases, y tends to decrease.
2. **Strength**:
    - The closer the value of r is to 1 or -1, the stronger the linear relationship.
    - The closer the value of r is to 0, the weaker the linear relationship.

Assumptions

1. **Linearity**: The relationship between x and y is linear.
2. **Homoscedasticity**: The variability of y is the same for all values of x.
3. **Normality**: Both x and y are normally distributed (or approximately so).

Understanding Pearson's r and its assumptions helps in properly interpreting the strength and direction of linear relationships in data.

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

Answer:

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.

WHY

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.

Types of Scaling

There are two common methods for scaling: normalization and standardization.

Normalized Scaling

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

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Pros:

- Keeps the original distribution of the data.
- All features are transformed to the same scale, which is beneficial for algorithms that assume feature comparability.

Cons:

- Sensitive to outliers, as outliers can significantly affect the minimum and maximum values.

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**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Pros:

- Reduces the effect of outliers because it centers the data around the mean.
- Useful for algorithms that assume the data is normally distributed.

Cons:

- Does not bound the data within a specific range, which may not be suitable for all applications.

Key Differences Between Normalization and Standardization

1. **Range**:
    - Normalization scales data to a fixed range, typically [0, 1].
    - Standardization transforms data to have a mean of 0 and a standard deviation of 1.

2. **Impact of Outliers**:
    - Normalization is more affected by outliers since it relies on the min and max values.
    - Standardization is less affected by outliers because it uses the mean and standard deviation.
3. **Use Cases**:
    - Normalization is useful when the algorithm does not make any assumptions about the distribution of the data and when all features need to be compared on the same scale.
    - Standardization is beneficial when the algorithm assumes normally distributed data, such as in Principal Component Analysis (PCA) or when dealing with data that has varying ranges and the presence of outliers.

Question 5. You might have observed that sometimes the value of VIF is infinite.

Why does this happen? (3 marks)

Answer :

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a set of predictor variables in a regression model. Multicollinearity occurs when one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy.

Calculation of VIF

For a given predictor XiX_iXi, the VIF is calculated as:

$$VIF_i = \frac{1}{1-R_i^2}$$

where R-squareis the coefficient of determination of a regression of Xi on all the other predictors.

Why the VIF Can Be Infinite

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.

Perfect Multicollinearity

Perfect multicollinearity occurs when one predictor variable is an exact linear combination of the others. In other words, there is a perfect linear relationship between one predictor and the other predictors. This makes it impossible to estimate

the unique effect of any one predictor on the dependent variable because changes in one predictor are perfectly matched by changes in another.

Addressing Infinite VIF

1. **Remove Perfectly Collinear Variables**: Identify and remove the variables that are causing perfect multicollinearity.
2. **Check for Data Issues**: Ensure that there are no coding errors, duplicate columns, or highly correlated predictors that need to be addressed.
3. **Principal Component Analysis (PCA)**: Transform the predictors into a set of linearly uncorrelated components.

Understanding why the VIF can become infinite and how to address this issue is crucial for creating robust regression models that provide reliable and interpretable results.

Question 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in

linear regression. (3 marks

Answer:

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

Construction of a Q-Q Plot

1. **Order the Data**: Sort the data in ascending order.
2. **Calculate Theoretical Quantiles**: Determine the theoretical quantiles from the chosen distribution (e.g., normal distribution).
3. **Plot the Points**: Plot the data quantiles on the y-axis against the theoretical quantiles on the x-axis.
4. **Interpret the Plot**: If the data follows the theoretical distribution, the points should approximately form a straight line.

Use and Importance in Linear Regression

In the context of linear regression, Q-Q plots are primarily used to assess the assumption of normality of the residuals. Normality of residuals is one of the key assumptions in linear regression, particularly for hypothesis testing and the construction of confidence intervals.

Assessing Normality of Residuals

1. **Residual Analysis**: After fitting a linear regression model, the residuals (the differences between observed and predicted values) should be normally distributed for valid inference.
2. **Creating a Q-Q Plot for Residuals**: Plot the quantiles of the residuals against the quantiles of a normal distribution. This helps to visually inspect whether the residuals follow a normal distribution.
3. **Interpreting the Q-Q Plot**:
    - **Straight Line**: If the points fall approximately along the straight line, the residuals are normally distributed.
    - **Deviations from Line**: Deviations from the straight line indicate departures from normality. Patterns in these deviations can suggest specific issues:
        - **S-shaped curve**: Indicates heavy tails (leptokurtic distribution).
        - **Inverted S-shaped curve**: Indicates light tails (platykurtic distribution).
        - **Other patterns**: Suggest skewness or other deviations from normality.

Importance in Linear Regression

1. **Validity of Statistical Tests**: Many statistical tests in regression (e.g., t-tests for coefficients, F-tests) assume normality of residuals. Q-Q plots help verify this assumption.
2. **Confidence Intervals**: The construction of accurate confidence intervals for regression coefficients relies on normally distributed residuals.
3. **Model Diagnostics**: Q-Q plots are a diagnostic tool to identify potential issues with the model, such as non-normality, outliers, or skewness.
4. **Improving Model**: If the residuals deviate significantly from normality, it may indicate the need for model improvement, such as transforming the response variable or adding/removing predictors.

Example of Q-Q Plot Interpretation

Suppose we fit a linear regression model to a dataset and obtain the residuals. We then create a Q-Q plot for these residuals:

1. **Straight Line**: The points closely follow the 45-degree line, suggesting the residuals are approximately normally distributed. This indicates that the normality assumption holds.
2. **Deviations from Line**: The points deviate systematically from the line, indicating that the residuals are not normally distributed. This could be due to skewness, heavy tails, or other issues.

By examining the Q-Q plot, we can make informed decisions about the adequacy of the regression model and whether further steps are needed to meet the assumptions of linear regression. This ensures more reliable and valid inferential statistics from the model.