**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)

Answer: 7 categorical variables are there in my dataset and I plot boxplot and found below insights

- Season: Fall season has more booking.
- Year: In the year 2019 bookings increased drastically compared to 2018
- Holiday: when it's not a holiday bookings are low.
- Weekday: Thu, Fri, Sat and Sun have more number of bookings as compared to the start of the week
- Working day: Working day has not much impact on bookings.
- Weather: When weather is clear, bookings are high.
- Month: Sept and Oct has high number of bookings. Start from Jan bookings increased till Sept,Oct then decreased.

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

Answer:

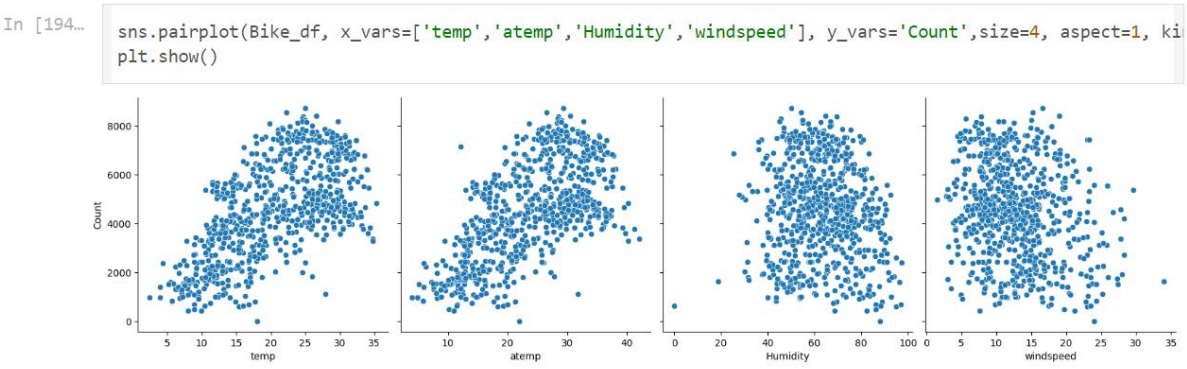
Drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 2 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not Holiday, then It is obvious Holiday. So we do not need 2nd variable to identify the not Holiday.

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

Answer:

From below picture, temp and atemp has high correlationwith Target variable Count.



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

Answer:

We can validate the assumptions of Linear Regression after building the model on the following training set by below method:

1) Fitted regression line is linear.

2) Error terms came out normally distributed with mean as 0

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:

From below picture,Temp,Year_2019,Season_winter features contributing significantly towards explaining the demand of the shared bikes

In [252…

```
param = pd.DataFrame(lr.params)
param.insert(0,'Variables',param.index)
param.rename(columns = {0:'Coefficient value'},inplace = True)
param['index'] = list(range(0,15))
param.set_index('index',inplace = True)
param.sort_values(by = 'Coefficient value',ascending = False,inplace = True)
param
```

Out[252…

| index | Variables | Coefficient value |
|---|---|---|
| 13 | temp | 0.472823 |
| 12 | Year_2019 | 0.234361 |
| 0 | const | 0.173663 |
| 3 | Season_winter | 0.079699 |

**General Subjective Questions**

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

Answer: Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

Mathematically the following equation,

$Y = m*X + b$

Where X = dependent variable (target)

Y = independent variable

m = slope of the line

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

Multi-collinearity –

o Linear regression model assumes that there is very little or no multi-collinearity in

the data. Basically, multi-collinearity occurs when the independent variables or

features have dependency in them.

Auto-correlation –

o Another assumption Linear regression model assumes is that there is very little or

no auto-correlation in the data. Basically, auto-correlation occurs when there is

dependency between residual errors.

Relationship between variables –

o Linear regression model assumes that the relationship between response and

feature variables must be linear.

Normality of error terms –

o Error terms should be normally distributed

Homoscedasticity –

o There should be no visible pattern in residual values.

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

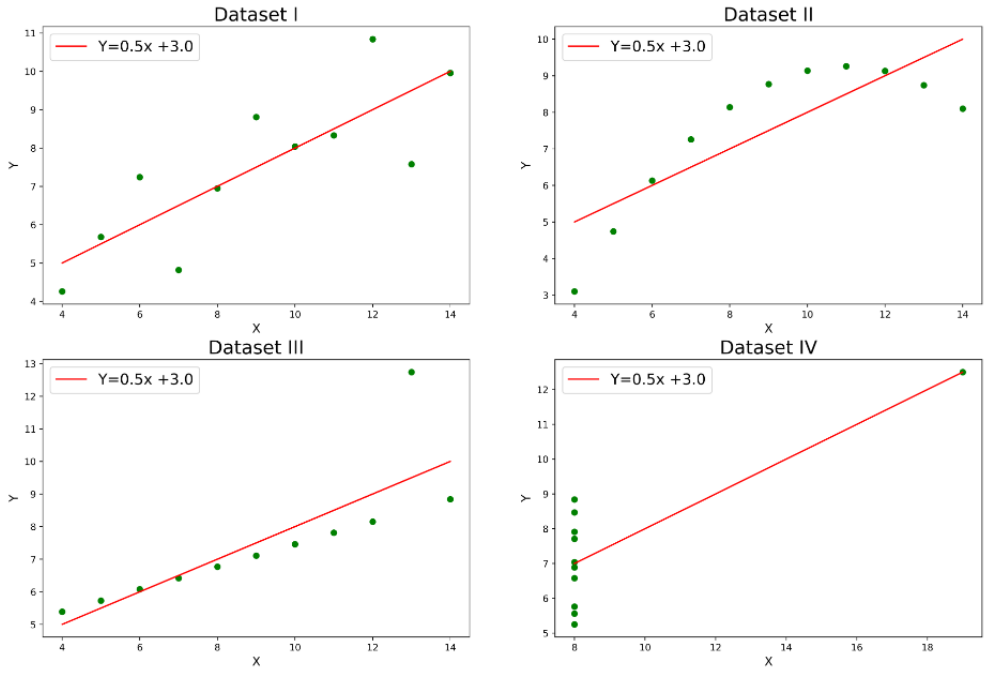
Answer:

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

**Anscombe's quartet** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's Quartet Dataset
The four datasets of 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 |




Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

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

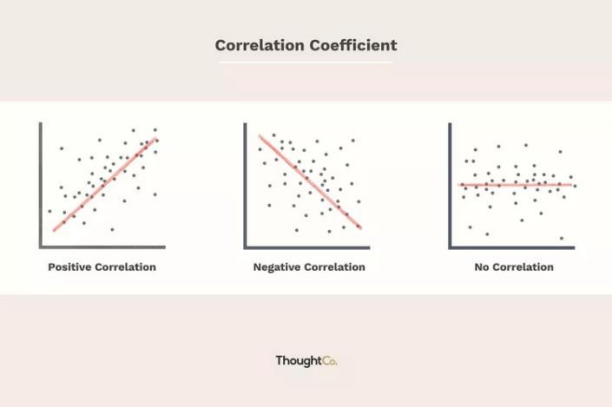
Answer:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

R Between 0 and 1 - When one variable changes, the other variable changes in the **same direction.**

R is 0 There is **no relationship** between the variables.

R Between 0 and –1 When one variable changes, the other variable changes in the **opposite direction.**




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.

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.

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

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).

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

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)

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a

correlation between the variables. If the VIF is 4, this means that the variance of the model

coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent

variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2)

infinity. To solve this 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)

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.

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.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain

some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests