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:
Significant variables in prediction of count of the number of boom bikes(cnt) are:
yr - positively correlated
temp - positively correlated
season (spring) - negatively correlated
mnth (july) - negatively correlated
mnth (September) - positively correlated
weathersit (good) - positively correlated
weathersit (bad) - negatively correlated

- Demand for boom bikes is expected to increase next year
- rainy season sees a decline in demand
- Demand increases when weather is favourable for bike riders and declines in harsh weather situations.

2. Why is it important to use drop_first=True during dummy variable creation?
Answer:
drop_first=True creates k-1 dummy variables for k categorical values. Not using it will create k dummy variables which is in-effect one-hot encoding.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer:
Following steps were taken to validate the assumptions of Linear Regression after building the model on training set:

- Residual Analysis : Histogram of error terms (normal curve)
- Scatterplot of y_test vs y_pred
- Comparing r-square of test and training data

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

Answer:

cnt = 2074.71 * yr + 3508.51 * temp - 1416.46 * season_spring - 645.69 * mnth_jul + 434.77 * mnth_sept - 1738.81 * weathersit_bad + 711.58 * weathersit_good

Top three features:

1) Temp
2) weathersit_bad
3) season_spring

General Subjective Questions

=======================

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a machine learning algorithm based on supervised learning. It executes a regression operation. Regression uses independent variables to model a goal prediction value. It is mostly used to determine how variables and forecasting relate to one another. Regression models vary according to the number of independent variables they use and the type of relationship they take into account between the dependent and independent variables.

The task of predicting a dependent variable's value (y) based on an independent variable is carried out using linear regression (x). Therefore, x (the input) and y (the output) are found to be linearly related by this regression technique (output). Thus, the term "linear regression" was coined.

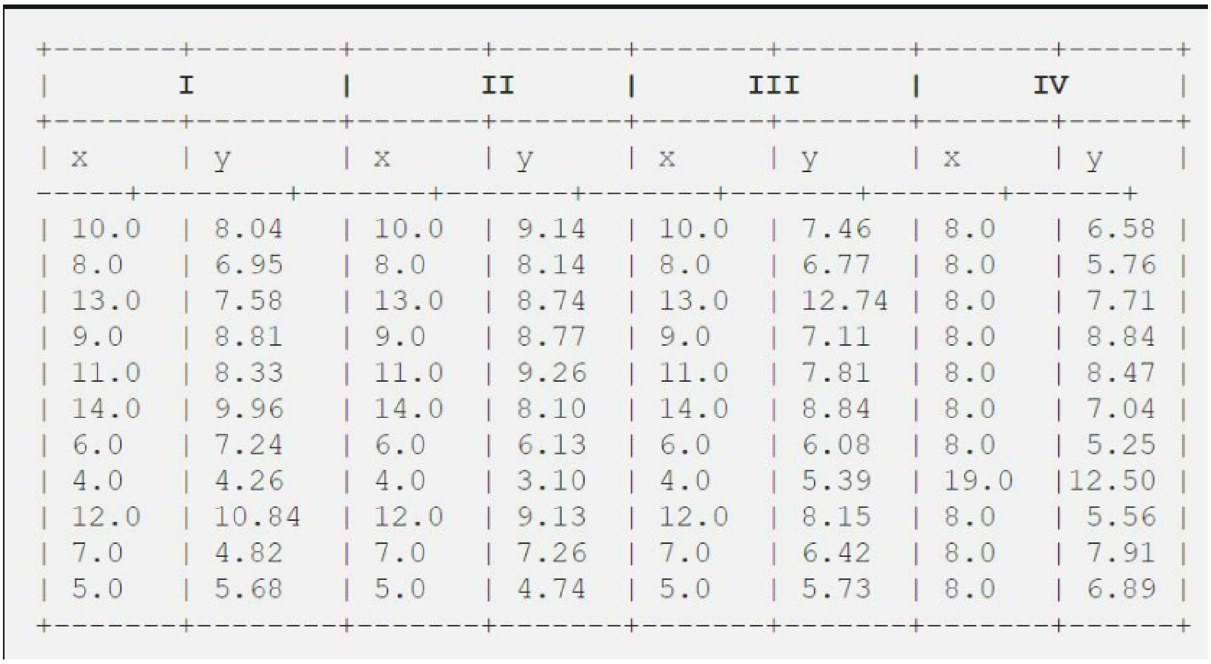
The regression line is the line that fits our model the best.

Linear regression is done on numerical variables.

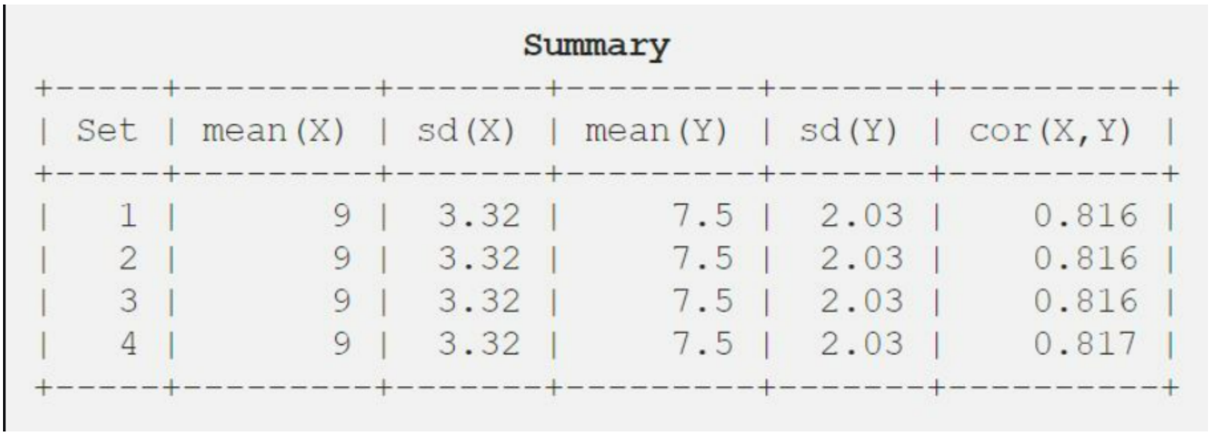
2. Explain the Anscombe’s quartet in detail.

Answer:

Anscombe's quartet consists of four datasets that, despite sharing a lot of basic statistical characteristics, look radically different when graphed. There are eleven (x,y) points per dataset. The statistician Francis Anscombe created them in 1973 to illustrate the value of charting data before analysing it as well as the impact of outliers on statistical features.

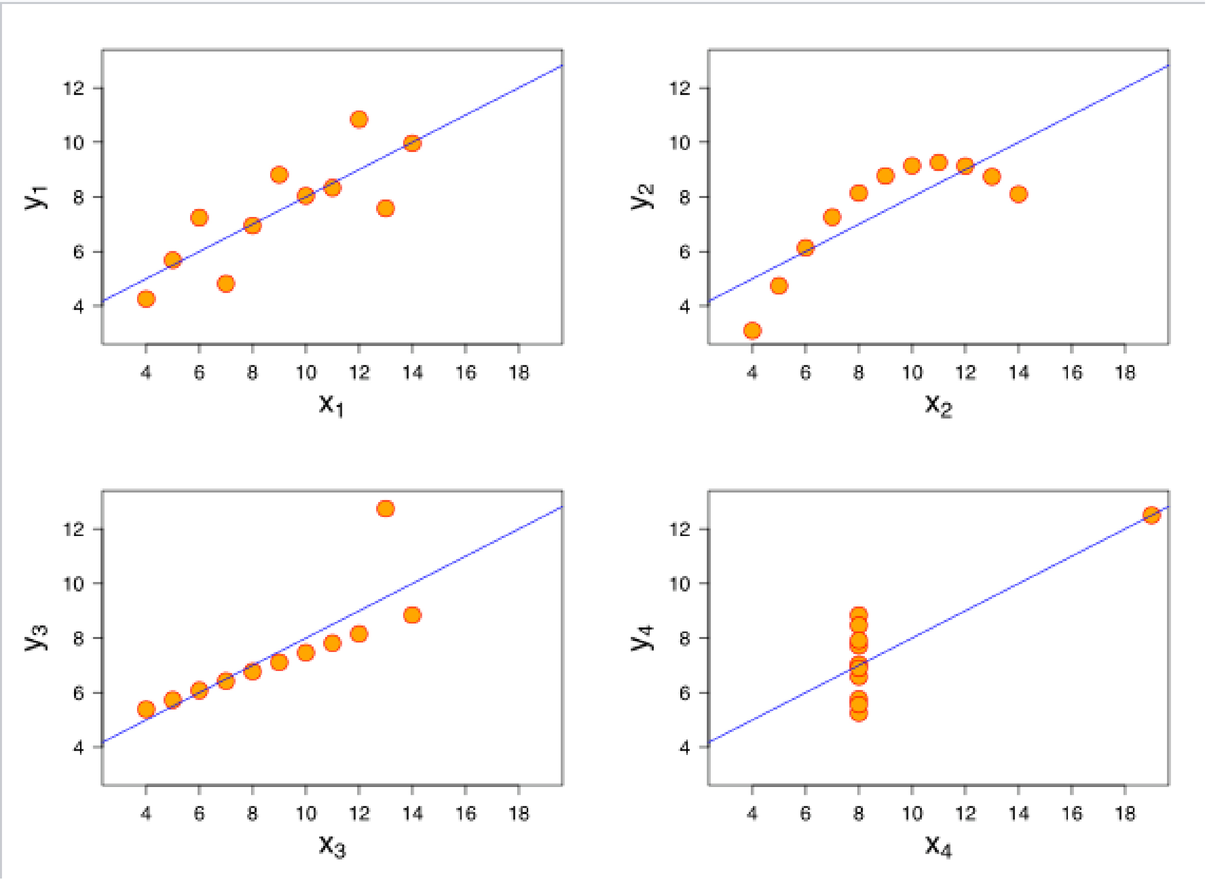


```
+-------+--------+-------+--------+-------+--------+-------+------+
|       I        |       II       |      III       |      IV       |
+-------+--------+-------+--------+-------+--------+-------+------+
| x     | y      | x     | y      | x     | y      | x     | y    |
-----+--------+-------+--------+-------+-------+-------+------+
| 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 |
+-------+--------+-------+--------+-------+--------+-------+------+
```



```
                           Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

As seen above, the datasets have the same mean, standard deviation and correlation but their graphical visualisation looks like this:



The quartet is still frequently used to highlight the need to visually examine a set of data before beginning an analysis based on a specific sort of relationship and the insufficiency of fundamental statistical features for characterising real-world datasets.

3. What is Pearson’s R?

Answer:

The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Answer:

Scaling helps in speeding up calculations of an algorithm.
Helps in creating better visualisations.

Normalised Scaling:

Scales values between [0, 1] or [-1, 1].
It is affected by outliers
It is used when features are of different scales
Formula:
X_new = (X - X_min)/(X_max - X_min)

Standardised Scaling:
It is not affected by outliers much.
It is not bound to a certain range.
It is used when we want to ensure zero mean and unit standard deviation
Formula:
X_new = (X - mean)/Std

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

Answer:

If the VIF is 3, it means that the variance of the model coefficient is inflated by a factor of 3 due to the presence of 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)

In other words, VIF = infinity if there is perfect correlation.

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

Answer:

When comparing the morphologies of two distributions, a Q-Q plot a.k.a quantile-quantile plot is used to show how characteristics like location, scale, and skewness are the same or different in the two distributions. Theoretical distributions or sets of data can be compared using Q-Q graphs. Q-Q plots can be thought of as a non-parametric method of comparing the underlying distributions of two samples of data. In general, a Q-Q plot is more effective than the widely used method of comparing the histograms of the two samples to accomplish this, but it requires more expertise to comprehend.

A Q–Q plot is a probability plot, which is a created by plotting the quantiles of two probability distributions against each other.