**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?**

* Bike rentals are higher in fall season.
* Bike rentals showing growth.
* Bike rentals are similar in all days of the week.

1. **Why is it important to use drop\_first=True during dummy variable creation?**

N-1 dummy variables are sufficient for N levels of categorical variables.

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

Registered column has highest correlation.

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

By checking if p value is less than 5% and VIF is less than 5.

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

Temperature, season and weather.

**General Subjective Questions**

1. **Explain the linear regression algorithm in detail.**

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.

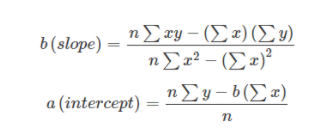
Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

Y = a +bx

Where a and b given by the formulas:

Here, x and y are two variables on the regression line.



b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

1. **Explain the Anscombe’s quartet in detail.**

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

**Data**

|  |  |  |
| --- | --- | --- |
| **Property** | **Value** | **Accuracy** |
| [Mean](https://en.wikipedia.org/wiki/Mean) of *x* | 9 | exact |
| Sample [variance](https://en.wikipedia.org/wiki/Variance) of *x* : sx{\displaystyle ^{2}} | 11 | exact |
| Mean of *y* | 7.50 | to 2 decimal places |
| Sample variance of *y* : sy{\displaystyle ^{2}} | 4.125 | ±0.003 |
| [Correlation](https://en.wikipedia.org/wiki/Correlation) between *x* and *y* | 0.816 | to 3 decimal places |
| [Linear regression](https://en.wikipedia.org/wiki/Linear_regression) line | *y* = 3.00 + 0.500*x* | to 2 and 3 decimal places, respectively |
| [Coefficient of determination](https://en.wikipedia.org/wiki/Coefficient_of_determination) of the linear regression : {\displaystyle R^{2}} | 0.67 | to 2 decimal places |

* The first [scatter plot](https://en.wikipedia.org/wiki/Scatter_plot) (top left) appears to be a simple [linear relationship](https://en.wikipedia.org/wiki/Linear_relationship), corresponding to two [variables](https://en.wikipedia.org/wiki/Variable_(mathematics)) correlated where y could be modelled as [gaussian](https://en.wikipedia.org/wiki/Normal_distribution) with mean linearly dependent on x.
* The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the [Pearson correlation coefficient](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient) is not relevant. A more general regression and the corresponding [coefficient of determination](https://en.wikipedia.org/wiki/Coefficient_of_determination) would be more appropriate.
* In the third graph (bottom left), the distribution is linear, but should have a different [regression line](https://en.wikipedia.org/wiki/Regression_line) (a [robust regression](https://en.wikipedia.org/wiki/Robust_regression) would have been called for). The calculated regression is offset by the one [outlier](https://en.wikipedia.org/wiki/Outlier) which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
* Finally, the fourth graph (bottom right) shows an example when one [high-leverage point](https://en.wikipedia.org/wiki/High-leverage_point) is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The *x* values are the same for the first three datasets

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |

It is not known how Anscombe created his datasets.[[7]](https://en.wikipedia.org/wiki/Anscombe%27s_quartet#cite_note-ChatterjeeFirat-7) Since its publication, several

1. **What is Pearson’s R?**

Pearson's correlation coefficient is the [covariance](https://en.wikipedia.org/wiki/Covariance) of the two variables divided by the product of their [standard deviations](https://en.wikipedia.org/wiki/Standard_deviations). The form of the definition involves a "product moment", that is, the mean (the first [moment](https://en.wikipedia.org/wiki/Moment_(mathematics)) about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

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

**What is 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.

**Why is scaling performed:**

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.[[1]](https://en.wikipedia.org/wiki/Feature_scaling#cite_note-1)

It's also important to apply feature scaling if regularization is used as part of the loss function (so that coefficients are penalized appropriately).

**Difference between normalized scaling and standardized scaling:**

* Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
* Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

VIF is infinite if there is perfect correlation.

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

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.

Uses and Importance:

a) It can be used with sample sizes

b) 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.

It is used to check following scenarios:

If two data sets —

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior.