1. **There will be data development and analysis support towards building predictive models, outline the steps for developing a statistical model.**

Step 1: Understanding the Business Objective (Financial Modelling/ Target Marketing/ Increase Customer experience - whether to increase sales/target new customers/reduce cost of operation.)

Step 2: Define Modeling Goals

Step 3: Select/Get Data

Step 4: Prepare Data (Initial data cleaning, defining variables, appending multiple datasets, provide summary reports)

Step 5: Analyze and Transform Variables. (Univariate, Multivariate Analysis)

Step 6: Model Selection and Develop Models (Linear Regression, Logistic Regression, Clustering techniques)

Step 7: Validate Models (Testing), Optimize and Profitability

-Document Methodology and Models

- Implement Models and UAT- Perform UAT to ensure model is implemented correctly.

-Monitoring and Performance Tracking- Monitor performance of model on periodic basis.

1. **List the critical assumptions for valid use of simple linear regression model.**

The critical assumptions for a valid use of simple linear regression model are-

* The two variables should be in a linear relationship.
* The level of error in the model is roughly the same as the value of explanatory variable which means that the variance of errors is constant, this property is called as homoscedasticity.
* All the variables in the data set should be multivariate normal that is linear combination of the random variables should have a normal distribution. (We can use a P-P plot to compare the distribution of the residuals against a normal distribution by displaying their respective cumulative probabilities.
* There should not be much multicollinearity in the data. If there is some amount of multicollinearity in the data, the best solution would be to remove the variables that have a high variance inflation factor.
* Looking out for outliers- as they may affect the slope of a regression line and can have a unduly large influence on findings. It is important to scatterplot your potential outliers.

1. **The table below is a sample of website visits for visitor 00001. Visit number column provides the tracking on visits to the site; visit\_depth shows the pages sequence on each visit; and date\_time shows the time stamp on the website activity.**
   1. Outline the steps on creating column QTR that assigns the quarter and year to the activity

* At first, we import python libraries- numpy as np and panda as pd
* Read the data using-data read function
  1. Outline the steps on creating a visit re-numbering so that pages of the same visit will have the same number and the sequence starts at 1 for the first visit each quarter.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **visitor\_id** | **visit\_number** | **visit\_depth** | **date\_time** | **Qtr** | **QTR\_VISIT\_SEQ** |
| **1** | 00001 | 5 | 2 | 01DEC2020 : | 2020Q4 | 4 |
| **2** | 00001 | 5 | 3 | 01DEC2020 : | 2020Q4 | 4 |
| **3** | 00001 | 5 | 4 | 01DEC2020 : | 2020Q4 | 4 |
| **4** | 00001 | 2 | 1 | 05OCT2020 : 10:32 AM | 2020Q4 | 1 |
| **5** | 00001 | 3 | 1 | 22NOV2020 : 8:08 PM | 2020Q4 | 2 |
| **6** | 00001 | 3 | 6 | 22NOV2020 : 8:09 PM | 2020Q4 | 2 |
| **7** | 00001 | 3 | 7 | 22NOV2020 : 8:10 PM | 2020Q4 | 2 |
| **8** | 00001 | 4 | 1 | 01DEC2020 : 9:56 AM | 2020Q4 | 3 |
| **9** | 00001 | 4 | 4 | 01DEC2020 : 9:56 AM | 2020Q4 | 3 |
| **10** | 00001 | 5 | 1 | 01DEC2020 : | 2020Q4 | 4 |
| **11** | 00001 | 6 | 1 | 01DEC2020 : 1:20 PM | 2020Q4 | 5 |
| **12** | 00001 | 6 | 2 | 01DEC2020 : 1:20 PM | 2020Q4 | 5 |
| **13** | 00001 | 5 | 5 | 01DEC2020 : | 2020Q4 | 4 |
| **14** | 00001 | 5 | 6 | 01DEC2020 : | 2020Q4 | 4 |
| **15** | 00001 | 5 | 7 | 01DEC2020 : | 2020Q4 | 4 |
| **16** | 00001 | 5 | 8 | 01DEC2020 : | 2020Q4 | 4 |
| **17** | 00001 | 15 | 1 | 11APR2021 : | 2021Q2 | 1 |
| **18** | 00001 | 15 | 2 | 11APR2021 : 10:20 AM | 2021Q2 | 1 |
| **19** | 00001 | 15 | 7 | 11APR2021 : 10:20 AM | 2021Q2 | 1 |
| **20** | 00001 | 16 | 1 | 11MAY2021 : 10:37 AM | 2021Q2 | 2 |
| **21** | 00001 | 16 | 9 | 11MAY2021 : 10:37 AM | 2021Q2 | 2 |
| **22** | 00001 | 16 | 10 | 11MAY2021 : 10:37 AM | 2021Q2 | 2 |
| **23** | 00001 | 17 | 1 | 20MAY2021 : 12:40 PM | 2021Q2 | 3 |
| **24** | 00001 | 17 | 28 | 20MAY2021 : 12:43 PM | 2021Q2 | 3 |
| **25** | 00001 | 12 | 1 | 04FEB2021 : 4:40 PM | 2021Q1 | 1 |
| **26** | 00001 | 12 | 20 | 04FEB2021 : 4:51 PM | 2021Q1 | 1 |
| **27** | 00001 | 13 | 1 | 05FEB2021 : 2:13 AM | 2021Q1 | 2 |
| **28** | 00001 | 14 | 1 | 05FEB2021 : 9:05 AM | 2021Q1 | 3 |

1. **Consider the marketing performance report for channels during the open enrollment in Region If you are to provide an analysis to help marketing partners decide on planning for marketing in the next open enrollment season. Also list additional information you think will be useful to have available for this analysis.**