The goal of this case study is to work towards building a model to predict the total number of claims a customer is going to file with the company.

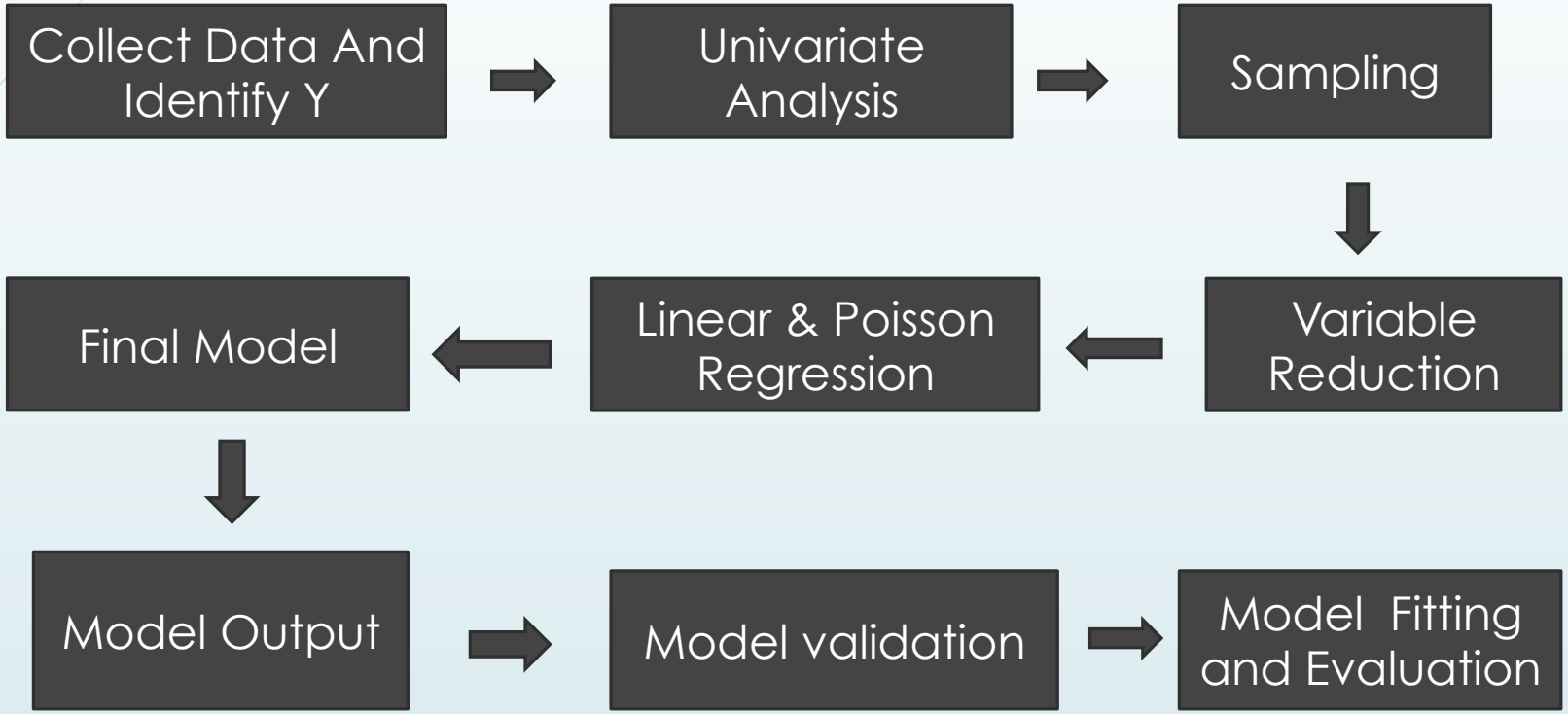
Problem Statement:

You are given a dataset containing policy information of motor insurance customers and the total claims they have filed with an insurance company. The goal of this case study is to work towards building a model to predict the total number of claims a customer is going to file with the company.

Objectives and Expected Output:

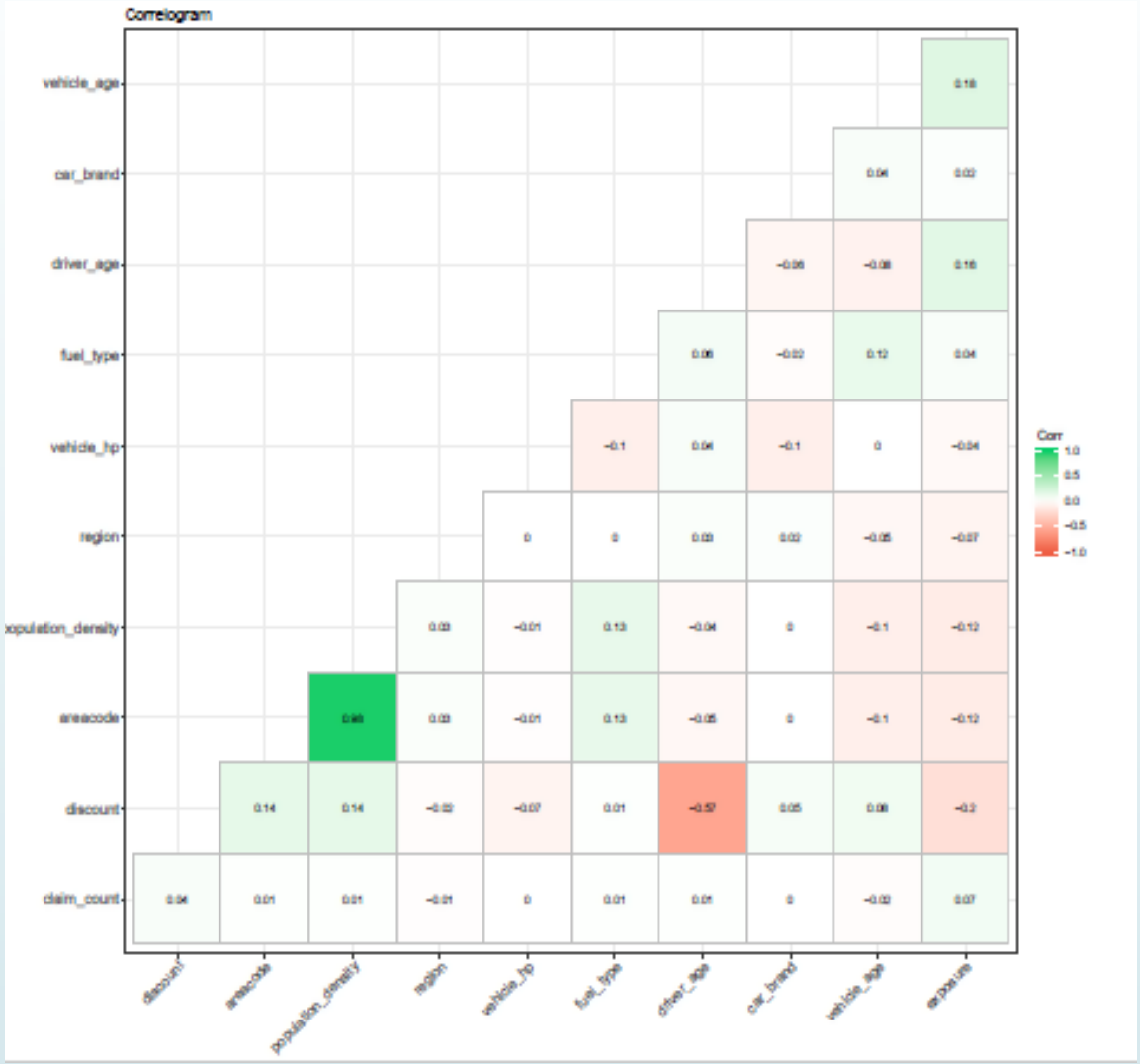
1. Build a model, using one or more techniques other than a GLM, to predict the frequency of claims (claim_count) a customer's going to make with the company.
2. In this case study, we are interested in all the steps involved in the model building process. Particularly data clean up, engineering new features, modelling and evaluation.
3. Log and report all the key decisions made along the process.
4. We expect a presentation of your data modelling process, model performance, and key decisions made along the way to the interview panel. This presentation should have no more than 5 slides.
5. Email the Presentation & all your source code. You are not expected to spend more than 3-6 hours on this exercise. Keep track of your time and let us know how much time you spent on this exercise.

Steps followed



Correlation Matrix

- Here the correlation matrix has been plotted using all the variables
- From the graph, It is seen that Area code and population density have highest correlation which is equals to 0.98
- Also driver age and discount have negative correlation which id equals to 0.57



Linear & Poisson Regression:

- **Linear Regression:**
  It is a way of finding a relationship between a single, continuous variable called Dependent or Target variable and one or more other variables (continuous or not) called Independent Variables.

  Linear Regression Equation:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots + b_kX_k$$

- b0 is the intercept the expected mean value of dependent variable (Y) when all independent variables (X's) are equal to 0. and b1is the slope. b1 represents the amount by which dependent variable (Y).

- **Poisson Regression:**
  Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters

  Poisson Regression Equation:

- Like in a linear regression model, we will model the conditional mean function using a linear combination $\beta\hat{\imath}t\ x{\downarrow}i$ of the explicative variables:

$$E[Y_i|x_i] = \exp(\beta^t x_i).$$

- Aim is then to estimate β , the unknown parameter in the model

Model Output:

Linear Regression

```
Call:
lm(formula = f2, data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2733 -0.0778 -0.0484 -0.0185 15.9653

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.980e-02  2.786e-03 -10.699  < 2e-16 ***
policy_desc       -1.906e-08  2.158e-10 -88.313  < 2e-16 ***
areacode           2.410e-03  3.133e-04   7.694 1.43e-14 ***
vehicle_age       -2.375e-03  6.328e-05 -37.535  < 2e-16 ***
discount           1.321e-03  2.564e-05  51.519  < 2e-16 ***
population_density 2.084e-07  1.083e-07   1.925 0.054284 .
exposure           4.427e-02  9.959e-04  44.454  < 2e-16 ***
fuel_type          2.350e-03  7.070e-04   3.324 0.000887 ***
driver_age         7.811e-04  2.821e-05  27.684  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2374 on 474325 degrees of freedom
Multiple R-squared:  0.02769,   Adjusted R-squared:  0.02768
F-statistic:  1689 on 8 and 474325 DF,  p-value: < 2.2e-16
```

Poisson Regression

```
Call:
glm(formula = f4, family = quasipoisson(link = "log"), data = trainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7714  -0.3591  -0.2692  -0.2018  13.0432

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -4.306e+00  4.845e-02 -88.875   <2e-16 ***
policy_desc       -3.880e-07  4.808e-09 -80.683   <2e-16 ***
areacode           5.324e-02  6.204e-03   8.582   <2e-16 ***
vehicle_age       -4.209e-02  1.357e-03 -31.020   <2e-16 ***
discount           2.053e-02  4.053e-04  50.659   <2e-16 ***
population_density 2.670e-06  2.001e-06   1.335    0.182
exposure           8.406e-01  2.024e-02  41.529   <2e-16 ***
driver_age         1.191e-02  5.046e-04  23.599   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.168589)

    Null deviance: 152320  on 474333  degrees of freedom
Residual deviance: 137973  on 474326  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Interpretation:

RMSE (Root Mean Square Error): It explains how close the actual data points are to the model's predicted values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

In the formula above, yi is the actual values of dependent variable and yi-hat is the predicted values, n - sample size.

- RMSE (Linear regression): 0.2370452 and RMSE (Poisson regression): 0.2351642
- as RMSE for Poisson regression is less than linear so final model will be Poisson regression

- **So Poisson regression model is best fitted model**

Thank You