# 1. EXECUTIVE SUMMARY

**Project Objective**

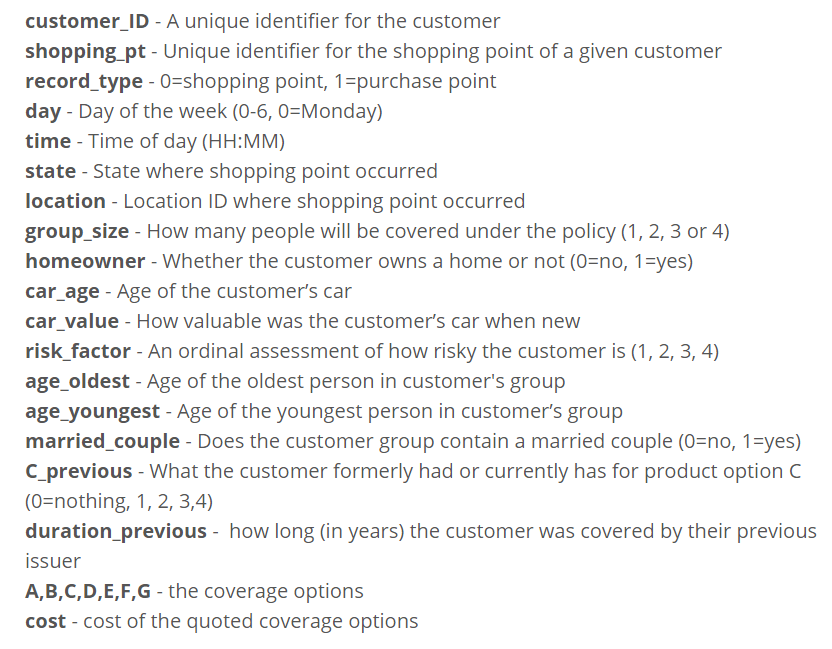
As a customer shops an insurance policy, he/she will receive several quotes with different coverage options before purchasing a plan. This is represented in this challenge as a series of rows that include a customer ID, information about the customer, information about the quoted policy, and the cost. Your task is to predict the purchased coverage options using a limited subset of the total interaction history. If the eventual purchase can be predicted sooner in the shopping window, the quoting process is shortened and the issuer is less likely to lose the customer's business. Keeping this in perspective, using the SEMMA approach, we tried to achieve the following objectives:

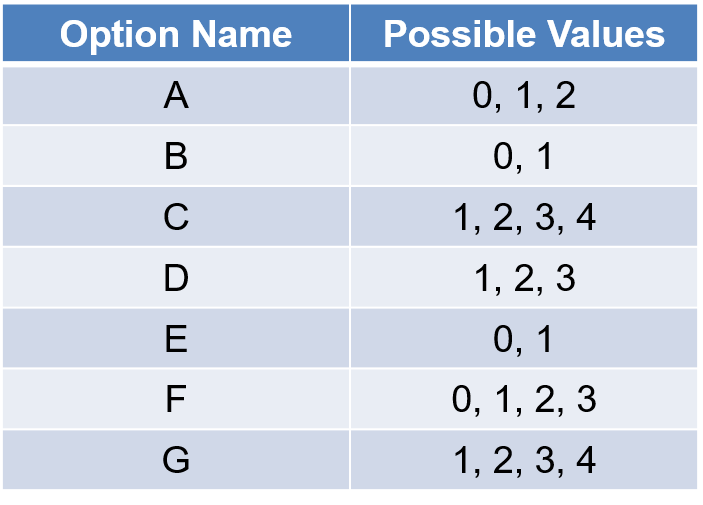
* Used last quote and individual characteristics to predict insurance coverage options
* Developed strategy for how to present optional coverages to customers to maximize company revenue

**Data Source**

The data is extracted from a kaggle competition. Following is the link for the dataset.

<https://www.kaggle.com/c/allstate-purchase-prediction-challenge>

**Data Definition**

There are 7 insurance options to buy form, which further have options to choose from. Thus, we must predict the exact combination of the insurance coverage options which is an outcome of 1800 unique combinations to be accurate in our prediction.

**Approach**For this project we followed the conventional SEMMA approach for the exploratory analysis, modifications, modeling, and assessing useful information for developing business strategies. Following is a holistic description of how we followed the SEMMA approach along with the technical steps performed in SAS JMP to achieve project objective.

* **SAMPLE**

The dataset consists of the historical purchase data for only 97k customers. So, we did not sample the dataset.

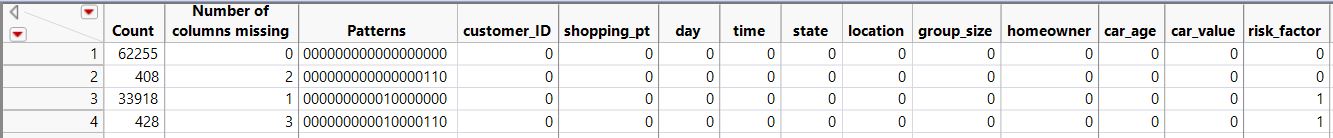
* **EXPLORE**

During the explore phase, our team worked on understanding the data definitions, identifying the continuous and categorical variables, and discovering relationships between the variables using the visualization techniques in JMP. We also identified the variables with missing values in the data-set and  
took the necessary decisions on how to deal with these missing values.

* **MODIFY**Based on the data exploration, the missing values in the variables was imputed. We tried to reduce data complexity and checked if majority of the information was being captured. A lot of new variables were created from the existing ones to cut down the curse of dimensionality, so that we don’t have to exclude important variables from the analysis.
* **MODEL**After preparing the data during the explore and modify phase, we focused on applying various modeling techniques like Logistic regression, Decision trees, Bootstrap forest, boosted forest, and neural networks to our data. Hence, multiple predictive models were built to search for an optimal combination of data that reliably predicted the desired outcome.
* **ASSESS**After we had all the different models, we compared the performances of these models with the baseline model to understand the value-added by the predictions. Then, depending upon various evaluation metrics we gauged the performance of the models and recommended the optimal models.

# 2. EXPLORATORY ANALYSIS

**Missing Data Analysis**

* Our team analyzed the data set for missing values using the “Missing Data Pattern” option in SAS JMP.
* Based on the Missing value exploration we found that risk\_factor was missing for 35.4% of values and C\_Previous and Duration\_previous were missing for 1% of the values.
* Figure 1 below shows the pattern for missing values in the data set.

**Figure 1**

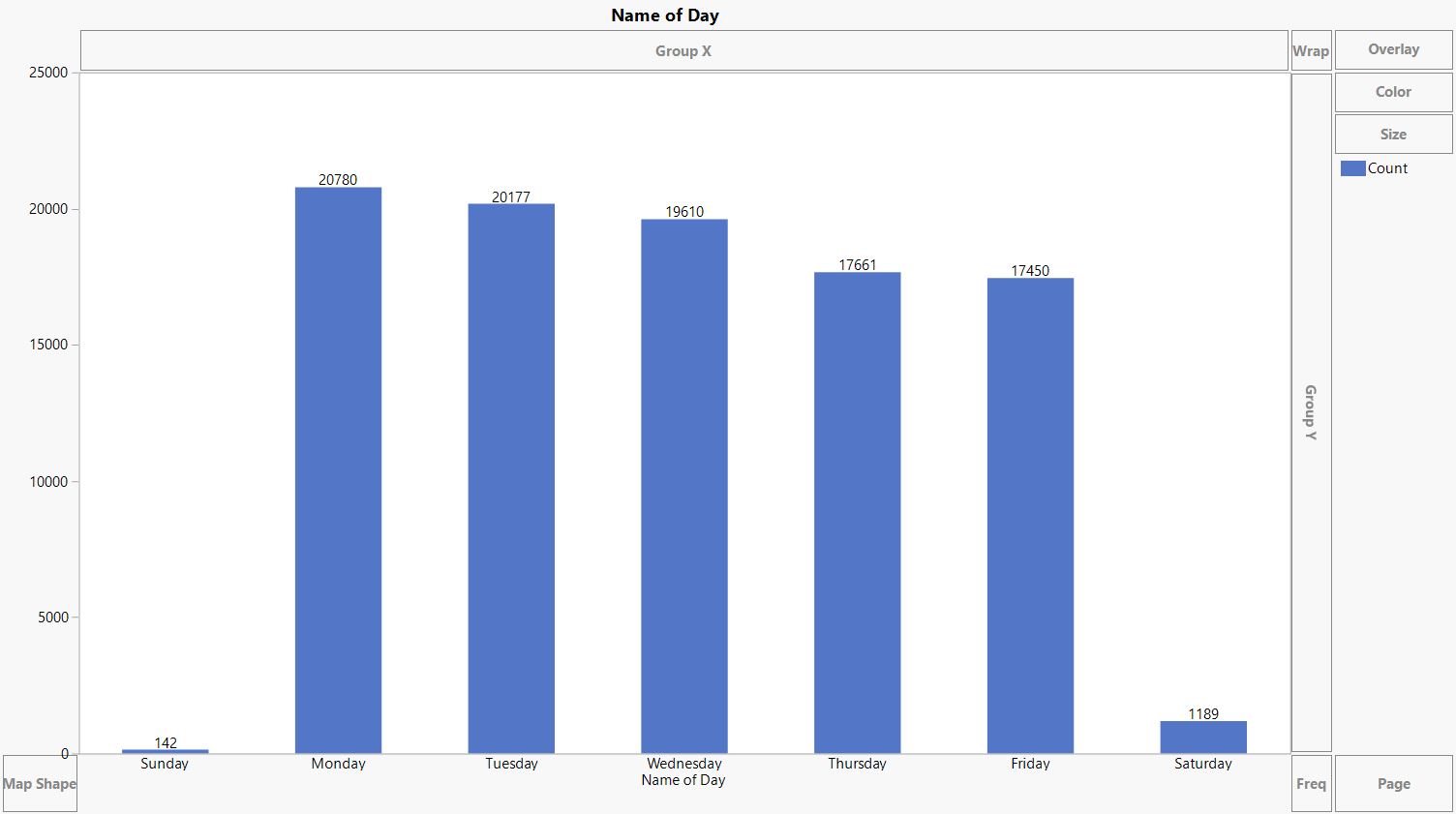
**Distributions and Key Observations**The AllState data set has a majority of Ordinal/nominal variables and a few continuous variables. Below is the list of the variables classified as Continuous/Ordinal/Nominal.

**Target Variable -** Option APurchase, Bpurchase, Cpurchase, Dpurchase, Epurchase, Fpurchase, Gpurchase

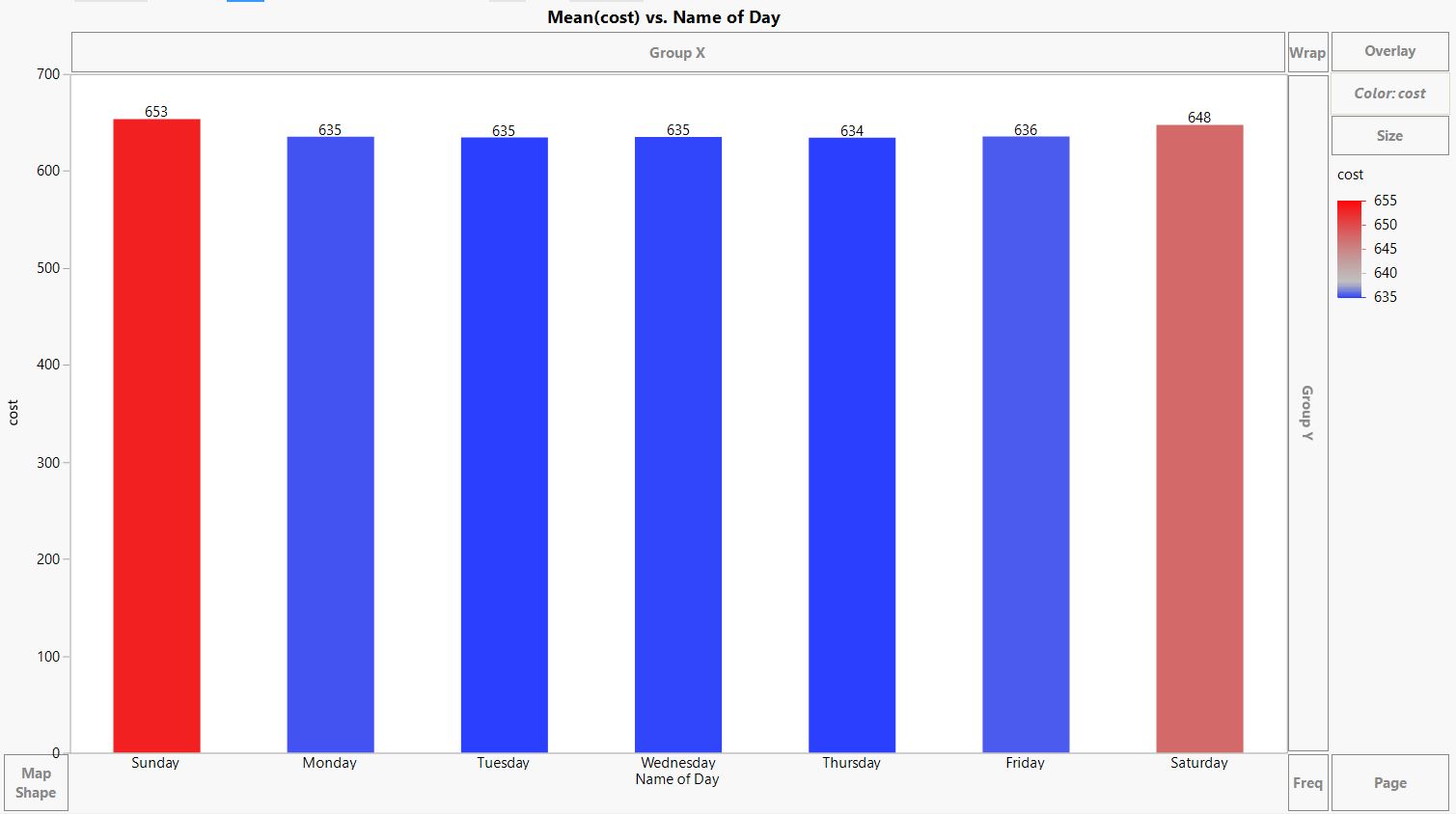
**Nominal variables -** state, homeowner, car\_value, married\_couple, C\_previous, changed\_from\_last\_quote, time\_of\_day, weekend/weekday, Last Quote A~G (before purchase)

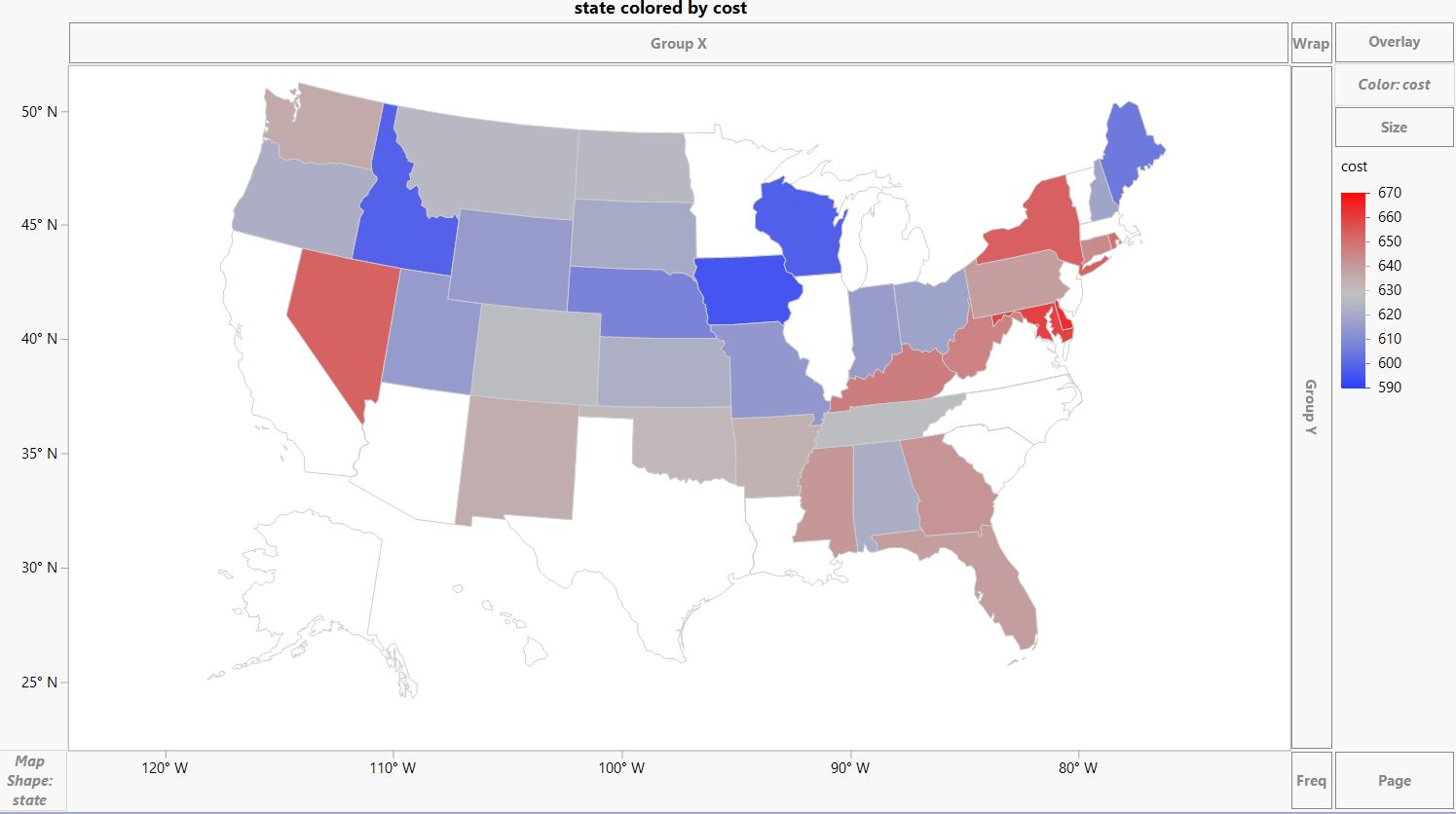
**Ordinal Variables -** imputed.risk.factor

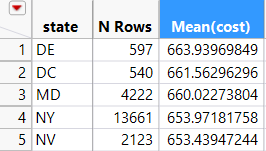
**Continuous Variables-** Shopping\_pt, group size, car\_age, age\_oldest, age\_youngest, duration\_previous, Cost, stability

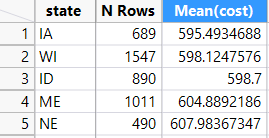
****As part of the data exploration phase, we conducted univariate analysis and multivariate analysis on all the variables. The objective was to explore how predictors are related with each other as well as with the target variables. Following are the descriptions for some of the conducted analyses.

**Figure 2**

**Figure 3**

Figure 2 shows the purchases made on each day of the week. We can clearly see that the purchases made on weekdays is way more than the purchases on weekends. But in figure 3, we can see that the average cost of the insurance package purchase on the weekends is more than those purchases made on weekdays. Hence the company can build strategy to drive more traffic on the weekends by holding some offers.

**Figure 4**

Figure 4 shows the relation between State and the average Cost. In the figure, the states shaded in red are the ones which have high average cost of packages, the states which are shaded in light red, grey and light blue are the ones which have medium average price for the packages and the states in dark blue have the lowest average cost of the packages. The first summary table on the right shows the top 5 states with highest average cost and the summary table at the bottom shows the 5 states with lowest average cost. Allstate insurance can design strategy to increase their presence in the states where the prices is low as there is a huge scope for profit.

**As there are very few continuous variables which are important contributors in our model, thus we did not study the Correlation matrix, Principal component analysis and outlier analysis for our project as there was no favorable outcome.**

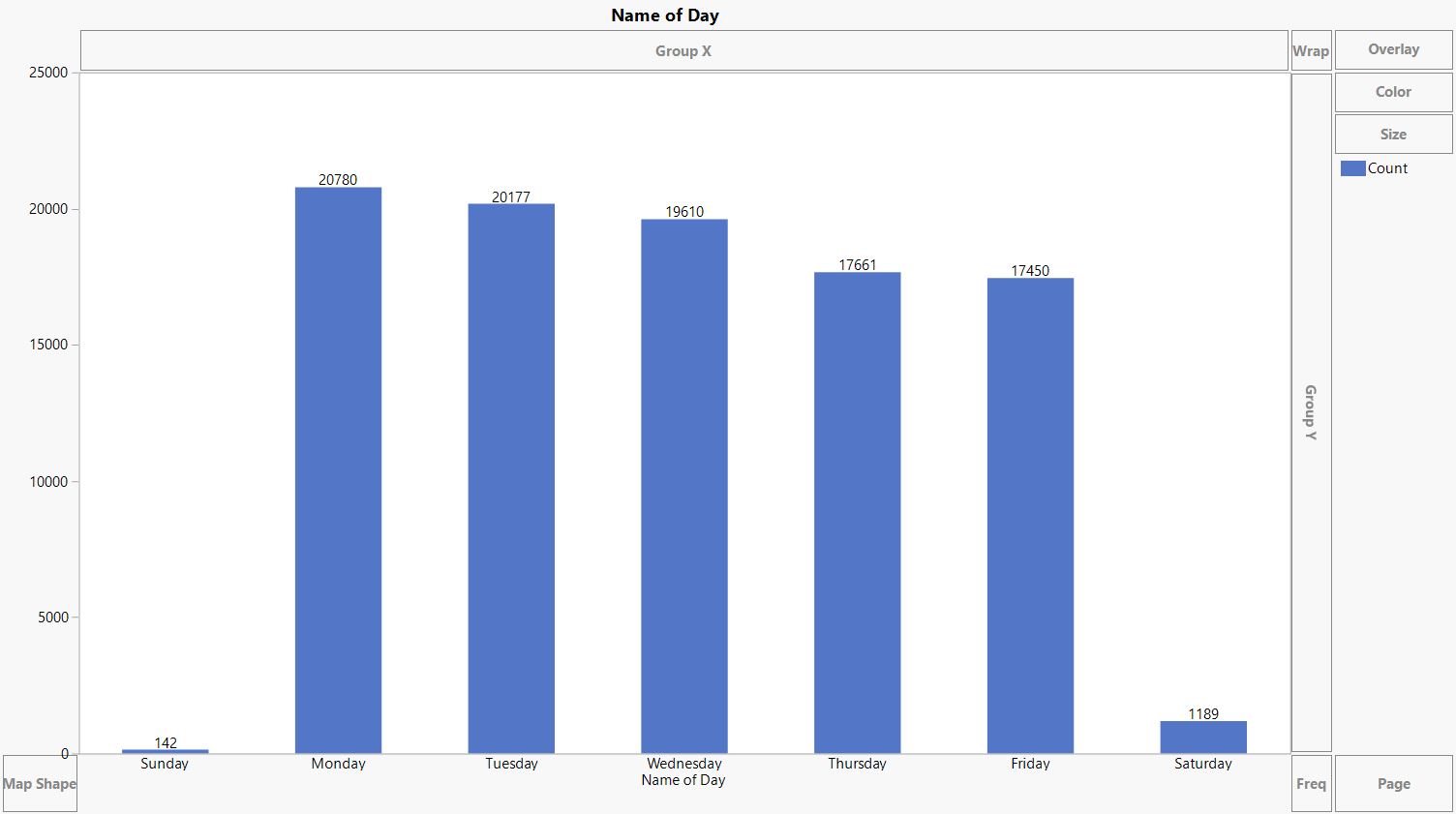
# 3. DATA MODIFICATIONS

1. Modification of data type of predictor variables

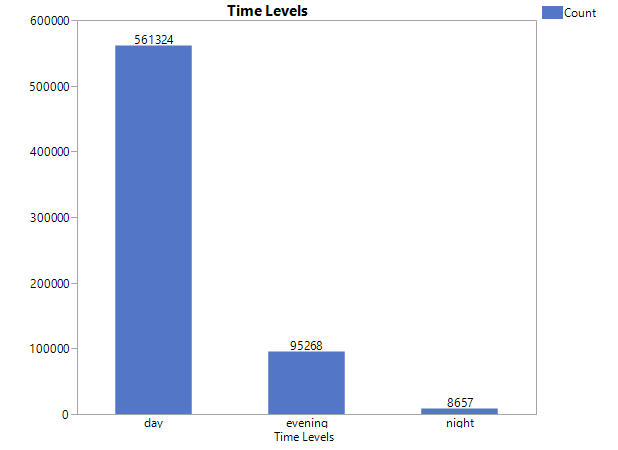
* Homeowner: continuous –> nominal
* married\_couple: continuous –> nominal
* C\_previous: continuous –> nominal
* A,B,C,D,E,F,G: continuous -> nominal
* record\_type: continuous -> nominal
* Risk\_factor: continuous -> ordinal

1. Grouping data of with multiple levels

* Day : (0-6) -> (Monday – Sunday) is divided into two levels Weekend/ Weekdays due to change in number of purchases at weekend compared to weekdays.



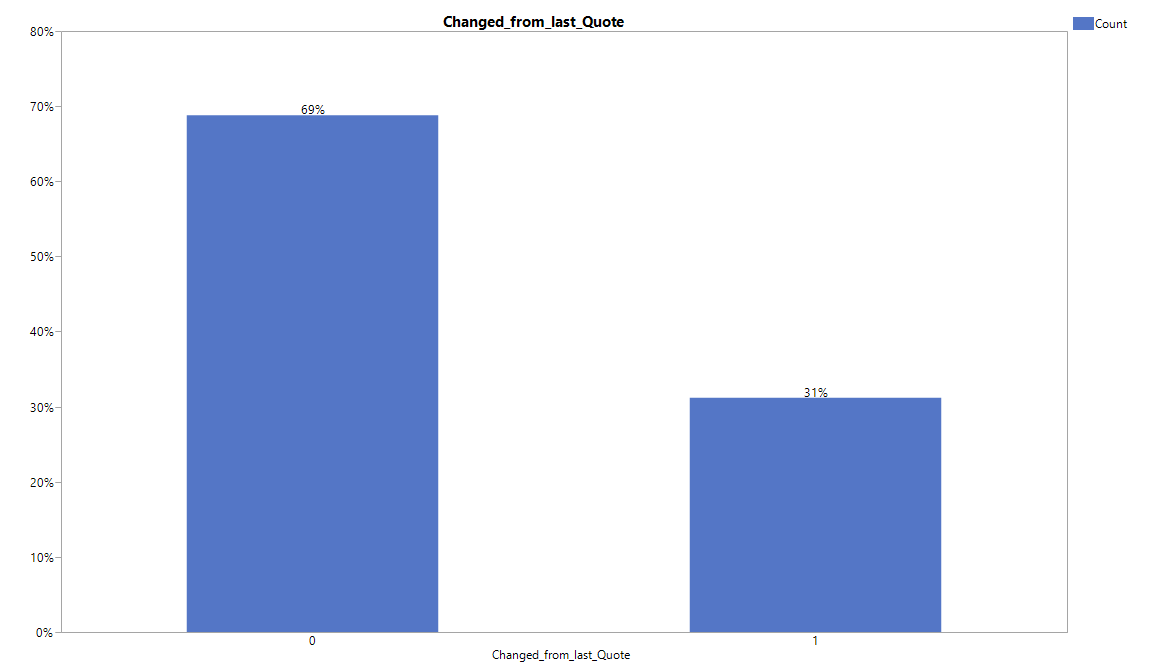
* Time : Time in 24 hours format is divided into three levels (day, evening, night) based on maximum variation in number of purchases in time span.



* Data Set divided into two section package with last shopping quote and final purchase.

Analysis of data shows that approx. 69% of the time customers did not change from the last quote.

This implies that last quote can be used as an important predictor variable for predicting the purchased package combinations.



1. Addition of new Predictor variables

New predictor variables are added as below:

**Time of Day**: To account for the change in level of time explained above

**Weekend/Weekday**: To account for change in levels of days as explained above

**Change\_From\_Last\_Quote**: to filter the data based on last quote

**Actual purchased**: To filter package combination of actual purchase

**Stability**: (no of quotes for single customer- no of unique quotes+1)/no of quotes for single customer

Stability specifies the how stable a person is at quotes while searching it before purchasing a final quote. This accounts the no of quotes that were missed in the data set when data set is filtered based on last quote and the Purchased quotes.

|  |  |  |
| --- | --- | --- |
| **Existing Variable** | **New variable** | **Formula** |
| **Time** | Hour | Substr(:time, 1, 2) , convert to continuous column |
| **Hour** | Time of the day | Match(:Hours continous, 6, "day", 7, "day", 8, "day", 9, "day", 10, "day", 11, "day", 12, "day", 13, "day", 14, "day", 15, "day", 16, "evening", 17, "evening", 18, "evening", "night") |
| **Day** | Name of Day | If(:Day== 0, "Monday", If(:Day == 1, "Tuesday", If(:Day == 2, "Wednesday", If(:Day == 3, "Thursday", If(:Day == 4, "Friday", If(:Day == 5,"Saturday","Sunday")))))) |
| **Day** | Weekday/Weekend | If(:Day<4, "Weekday","Weekend") |
| **risk\_factor** | Imputed.Risk.Value | Imputed using Linear Regression model for missing values |
|  | Actual purchased | If(:record\_type==1,1,0) |
|  | Changed\_from\_Last quote | If(:record\_type[Row() + 1] == 1, 1, 0) |
|  | Stability | (no of quotes for single customer- no of unique quotes+1)/no of quotes for single customer |

1. Modified response and predictor variable

**Response variables:** *APurchase, Bpurchase, Cpurchase, Dpurchase, Epurchase, Fpurchase, Gpurchase*

* **Predictor Variables:** *Shopping Point, Group Size, Home Owner, Car Age, Car Value, Age Oldest, Age Youngest, Married, Duration\_Previous, C\_Previous, Cost, Risk\_Factor(Imputed), Stability, Weekend/Weekdays, Last Quote A~G,* *Time of the day*