**1**

**Life Insurance sales – capstone**

**Project Notes – 1**

**SARATH KUMAR V**

**2**

**Table of Contents**

**1. Problem Statement: Life Insurance Data ......................**...................................................................................... **3 2. Need for this Study/Project**.................................................................................................................................................3 **3. Why is this (agent bonus) important for the business/company?** ....................................................3 **4. Data Report/Dictionary**..........................................................................................................................................................4 **5. Performing Exploratory Data Analysis (EDA)**……………………….………………………………………..4 **6. Checking for Unique Categorical Values**………………………………………………………………7 **7. Univariate/Bivariate Analysis**………………………………………………………………………………8 **8. Categorical Variable’s Univariate Analysis**…………………………………………………………13 **9. Bivariate Analysis ( Pairplot/Heatmap)**……………………………………………………………..18

**Table of Figures**

**Figure 1 – Univariate Analysis – Numerical Columns** ....................................................................... 8 **Figure 2 - Univariate Analysis – Categorical Columns** .....................................................................13 **Figure 3 - Categorical variable w.r.t Target Variable (Bivariate)**………………………………..15 **Figure 4 - Bivariate Analysis- Numerical Data Pairplot**...................................................................18 **Figure 5 - Bivariate Analysis- Numerical Data Correlation Heatmap**......................................19

**3**



**Problem Statement: Life Insurance Data**

· The dataset belongs to a leading life insurance company.

· The company wants to predict the bonus for its agents so that it may design appropriate engagement activity for their high performing agents and upskill programs for low performing agents.

**Need for this Study/Project**

· With this problem we want to better understand how the insurance company agents are performing, it’s not to underpay or overpay, as the payment is regulated by IRDA.

· With the predictions it’s better for the company to understand where they need to focus more as for agents selling less policies the company needs some booster training performs. As the policies are as good as the agents portray it to be to the potential customer.

· While the agents performing good i.e. selling more policies there needs to be a way to reward them, to make their contribution known so that they perform the same and even better in future.

**Why is this (agent bonus) important for the business/company?**

· A company is as good as their employers.

· For a Life Insurance Company, their agents are the best way to make the companies policies, aims, and perks known to the customer. Once the customer is intrigued by the policy delivery by the agent, its easier to convince the customer hence improving the sales and thereby motivating the agent as well.

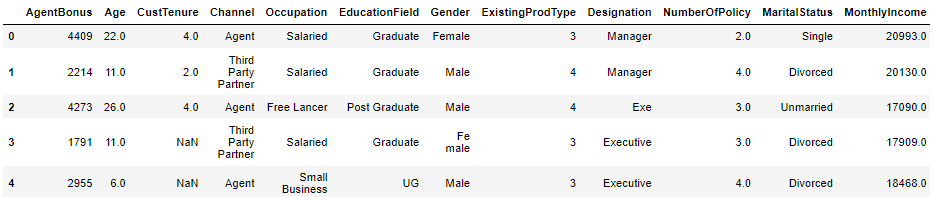
· With this, the market share of the company will gain more ground dominating the potential opponents.

· Moreover, the agents can be classified into categories giving the company better insight where the need to put more effort.

· The customer feedback can help the company develop improved and updated policies/products. Meeting customer needs.

· Hereby, the easiest way to retain their agents.

· Overall, multiplying and adding to company’s profit.

**4** **Data Report/Dictionary**



The following data is provided by Great Learning cover the Life Insurance Sales made by the company, the data dictionary consists of:

Variable CustID AgentBonus Age CustTenure Channel Occupation EducationField Gender

ExistingProdType Designation NumberOfPolicy MaritialStatus MonthlyIncome Complaint

ExistingPolicyTenure SumAssured

Zone

PaymentMethod

LastMonthCalls

CustCareScore

Description Unique customer ID

Bonus amount given to each agents in last month. Age of customer

Tenure of customer in organization.

Channel through which acquisition of customer is done. Occupation of customer

Field of education of customer Gender of customer

Existing product type of customer Designation of customer in their organization Total number of existing policy of a customer Marital status of customer

Gross monthly income of customer

Indicator of complaint registered in last one month by customer

Max tenure in all existing policies of customer

Max of sum assured in all existing policies of customer Customer belongs to which zone in India. Like East, West, North and South

Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly

Total calls attempted by company to a customer for cross sell

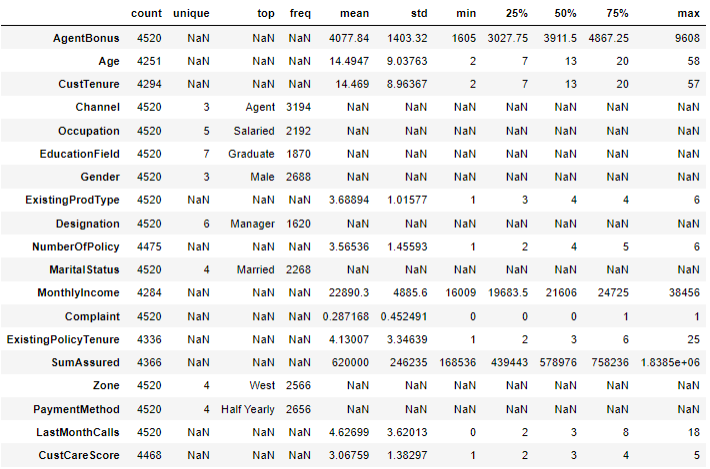
Customer satisfaction score given by customer in previous service call

**Performing Exploratory Data Analysis (EDA).**

**Head of the Data**

· I’ve removed CustID as it is irrelevant to agent bonus.

· Head gives us the idea of what the basic dataset looks like. · Complete list of all variables is not presented.

**5** **Shape of the dataset**



Total rows in the dataset: 4520 Total columns in the dataset: 19

**Descriptive Statistics of the Columns**

· The table includes the complete description for all variable with categorical variables included.

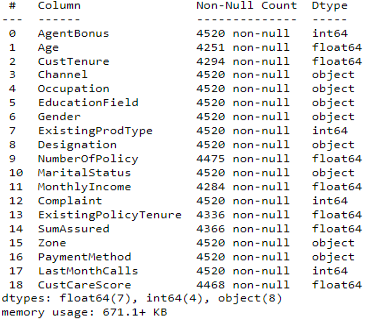
· The description includes, variable count, unique values, top frequently occurring categories like Agent-3194, mean, standard deviation, minimum, 25%, 50%(median), 75%, and maximum values present in the respective variables.

· Hence the ‘NaN’ here is observed for Categorical Variables as a string object cannot have numeric values.

· This we will change by encoding the data in future if needed.

· We can also observe the missing values as the count is not constant for all the variables. · The unique is only present for categorical variables which hold a specific category

· Example: Gender has male and female hence it should hold unique value of 2 but later we observed some subcategories needs to be renamed.

**6** **Info of the parameters**



· We have 7 parameters having ‘float’ data type.

· We have 4 parameters having ‘integer’ data type. · We have 8 parameters having ‘object’ data type.

· Age is shown as float, however we will later observe is its needed to change it to int or not, it won’t make any difference in our observations.

· We can clearly observe some missing values.

· Further count of missing values is provided below.

· CustID 0 · AgentBonus 0

· Age 269 · CustTenure 226 · Channel 0

· Occupation 0 · EducationField 0 · Gender 0 · ExistingProdType 0 · Designation 0 · NumberOfPolicy 45 · MaritalStatus 0

· MonthlyIncome 236 · Complaint 0

· ExistingPolicyTenure 184 · SumAssured 154 · Zone 0

· PaymentMethod 0 · LastMonthCalls 0 · CustCareScore 52

· **Number of duplicate rows = 0**

· The **Missing values** can affect the prediction’s hence need to be treated, hence the missing values are imputed with the **median values** in the respective column.

**7** **Checking for Unique Categorical Values.**



CHANNEL has 3 Unique Values. Online 468 Third Party Partner 858 Agent 3194 Name: Channel, dtype: int64

OCCUPATION has 5 Unique Values. Free Lancer 2

Laarge Business 153 Large Business 255 Small Business 1918 Salaried 2192 Name: Occupation, dtype: int64

EDUCATIONFIELD has 7 Unique Values. MBA 74

UG 230 Post Graduate 252 Engineer 408 Diploma 496 Under Graduate 1190 Graduate 1870 Name: EducationField, dtype: int64

GENDER has 3 Unique Values. Fe male 325 Female 1507 Male 2688 Name: Gender, dtype: int64

DESIGNATION has 6 Unique Values. Exe 127

VP 226 AVP 336 Senior Manager 676 Executive 1535 Manager 1620 Name: Designation, dtype: int64

MARITALSTATUS has 4 Unique Values. Unmarried 194

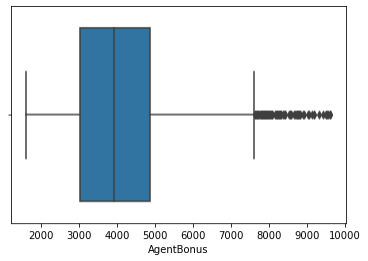
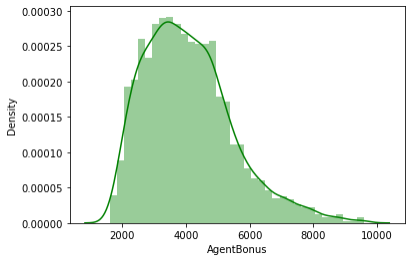
Divorced 804 Single 1254 Married 2268 Name: MaritalStatus, dtype: int64

ZONE has 4 Unique Values. South 6 East 64 North 1884 West 2566 Name: Zone, dtype: int64

PAYMENTMETHOD has 4 Unique Values. Quarterly 76

Monthly 354 Yearly 1434 Half Yearly 2656

**8**



· Here it can be observed that subcategories highlighted with a different colour shows an error in naming convention hence have to be renamed.

· Example: ‘Laarge’ and ‘Large’ Business can be put in the same category, the same for ‘UG’ and ‘Under Graduate’, ‘Graduate’ and ‘Post Graduate’, ‘Fe male’ and ‘Female’, and

‘Exe’ and ‘Executive’**.**

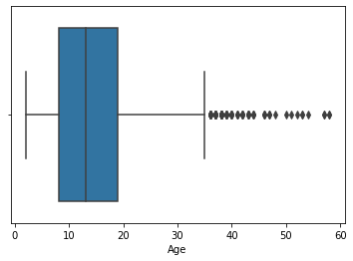
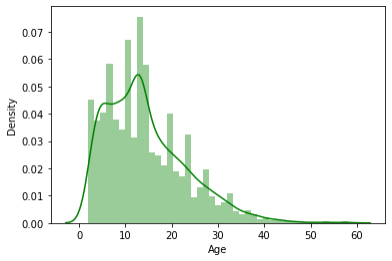
**Univariate/Bivariate Analysis**

**AgentBonus**

Figure 1(a) Distplot/Histplot - AgentBonus

· **The distribution of "AgentBonus" seems to be positively/right skewed.** · **The data ranges from 1605 to 9600.**

· **The box plot holds many outliers.**



**Age:**

Figure 1(b) Distplot/Histplot - Age

· **The distribution of "Age" seems to be positively/right skewed.** · **The data ranges from 2 to 58.**

· **The box plot holds many outliers.**

**9** **CustTenure:**

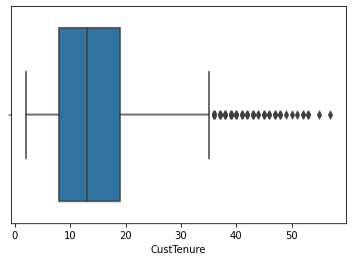
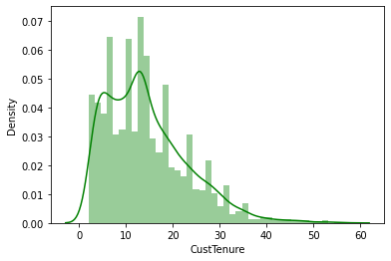
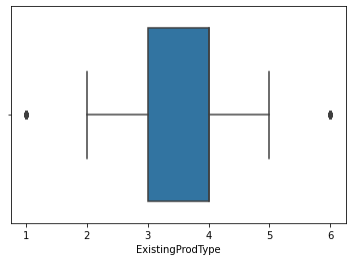
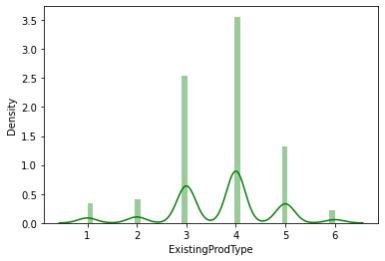


Figure 1(c) Distplot/Histplot - CustTenure

· **The distribution of "CustTenure" seems to be positively/right skewed.** · **The data ranges from 2 to 57.**

· **The box plot holds many outliers.**

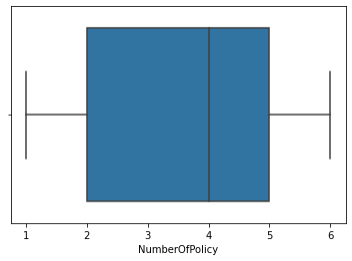
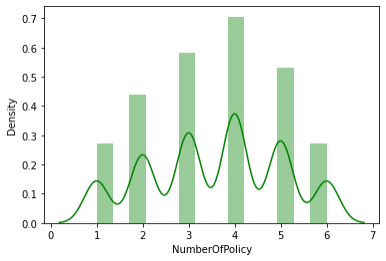


**ExistingProdType:**

Figure 1(d) Distplot/Histplot - ExistingProdType

· **The distribution of "ExistingProdType" seems to be slightly left skewed.** · **The data ranges from 1 to 6.**

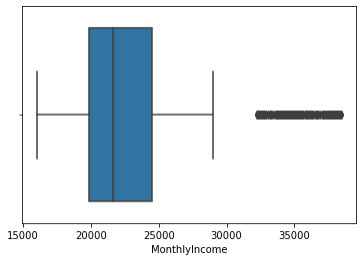
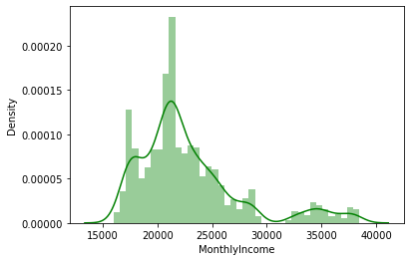
· **The box plot holds outliers.**



**NumberOfPolicy**

Figure 1(e) Distplot/Histplot - NumberofPolicy

**10**



· **The distribution of "NumberOfPolicy" seems to be slightly left skewed.** · **The data ranges from 1 to 6.**

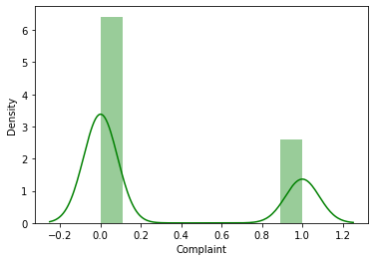
· **The box plot has no outliers.**

**MonthlyIncome**

Figure 1(f) Distplot/Histplot - MonthlyIncome

· **The distribution of "MonthlyIncome" seems to be positively/right skewed.** · **The data ranges from 16000 to 38500.**

· **The box plot holds many outliers**.



**Complaint**

Figure 1(g) Distplot/Histplot - Complaint

· **The distribution of "Complaint" seems to be positively/right skewed.** · **The data ranges from 0 to 1.**

· **The box plot holds no outliers.**

**11** **ExistingPolicyTenure**

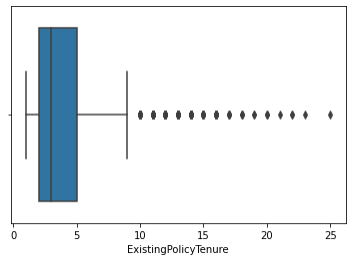
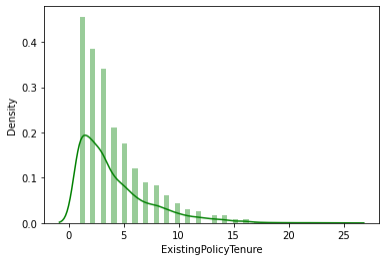
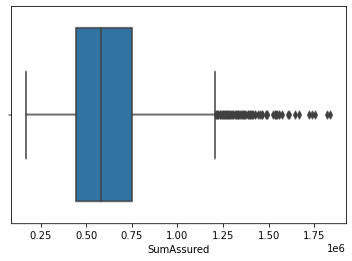
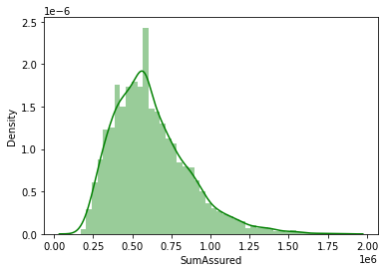


Figure 1(h) Distplot/Histplot - ExistingPolicyTenure

· **The distribution of "ExistingPolicyTenure" seems to be positively/right skewed.** · **The data ranges from 1 to 25.**

· **The box plot holds many outliers.**

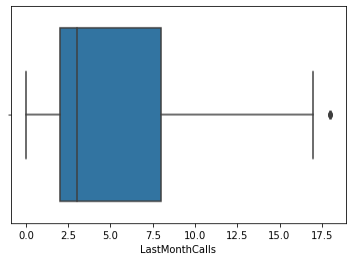
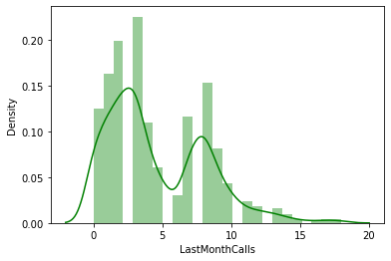


**SumAssured**

Figure 1(i) Distplot/Histplot - SumAssured

· **The distribution of "SumAssured" seems to be positively/right skewed.** · **The data ranges from 1.68 \* 105 to 1.83 \* 105.**

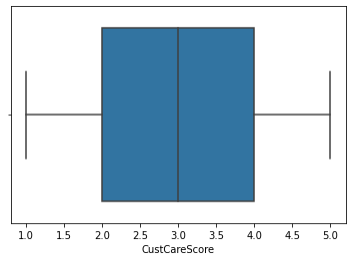
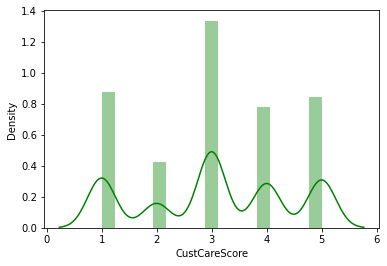
· **The box plot holds many outliers.**



**LastMonthCalls**

Figure 1(j) Distplot/Histplot - LastMonthCalls

**12**



· **The distribution of "LastMonthCalls" seems to be positively/right skewed.** · **The data ranges from 0 to 18.**

· **The box plot holds outliers.**

**CustCareScore**

Figure 1(k) Distplot/Histplot – CustCareScore

· **The distribution of "CustCareScore" seems to be slightly left skewed.** · **The data ranges from 1 to 5.**

· **The box plot holds no outliers**

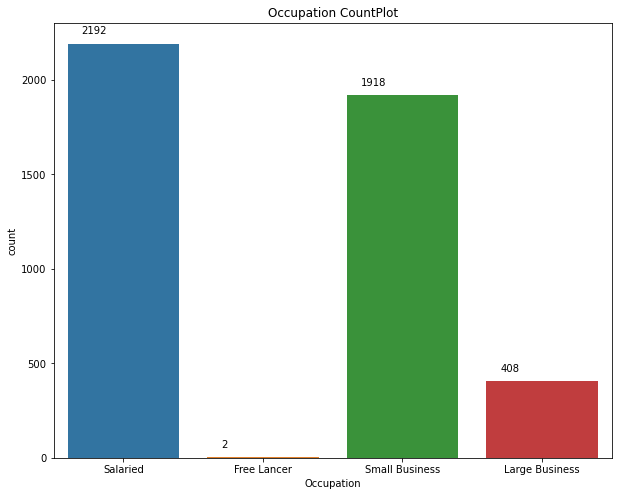
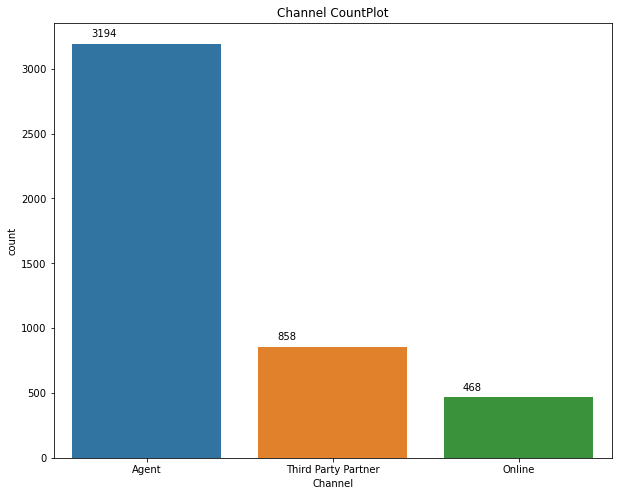
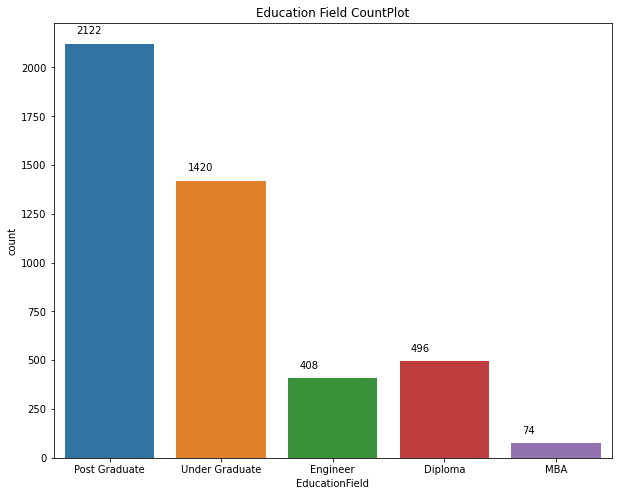
**Skewness**

AgentBonus 0.822348 Age 0.998425 CustTenure 0.981002 ExistingProdType -0.401100 NumberOfPolicy -0.108161 MonthlyIncome 1.434315 Complaint 0.941129 ExistingPolicyTenure 1.601730 SumAssured 1.002018 LastMonthCalls 0.810417 CustCareScore -0.138120

· We can observe skewness in the data with ExistingProdType, NumberofPoilicy and CustCareScore being negatively skewed.

· Rest all other parameters holds positive skewness the max being for ExistingPolicyTenure.

**13** **Categorical Variable’s Univariate Analysis**



**Education Field**

Post Graduate 0.47 Under Graduate 0.31 Diploma 0.11 Engineer 0.09 MBA 0.02

Most Customers approached are Post Graduates having 47% weightage.

Figure 2(a) Count Plot - EducationField

**Channel**

Agent 0.71

Third Party Partner 0.19 Online 0.10

Acquisition of a customer is mostly done Via an Agent having 71% weightage.

Figure 2(b) Count Plot - Channel

**Occupation**

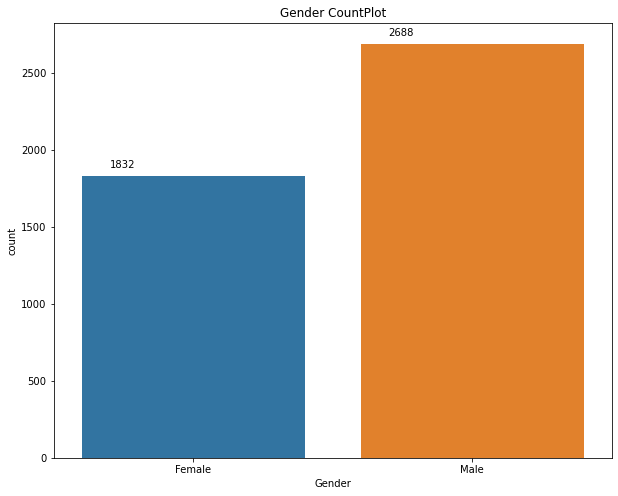
Salaried 0.48 Small Business 0.42 Large Business 0.09 Free Lancer 0.00

Most customers have Salaried Occupations Around 48%.

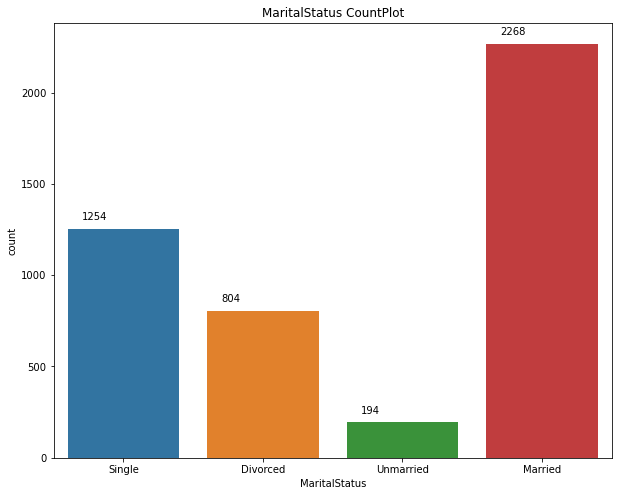
Here freelancers have a minute weightage**.**

Figure 2(c) Count Plot - Occupation

**14**



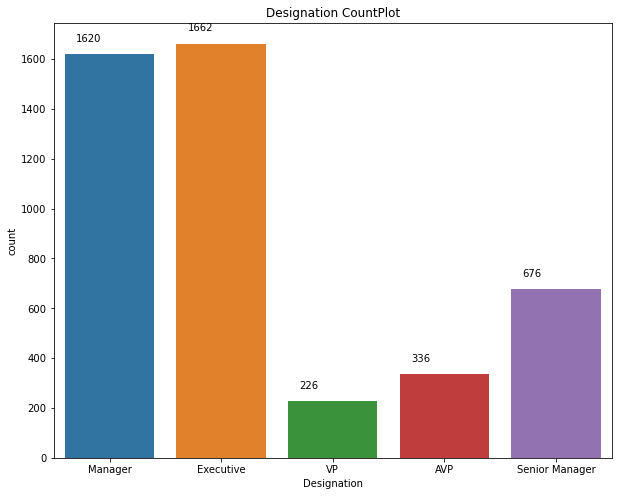
**Gender**



Male 0.59 Female 0.41

Approximately 59% of customers Are males.

Figure 2(d) Count Plot - Gender

**Designation**

Executive 0.37 Manager 0.36 Senior Manager 0.15 AVP 0.07

VP 0.05

Most customers are either a Executive or Managers having Weightage of 37% and 36% Respectively.

Figure 2(e) Count Plot - Designation

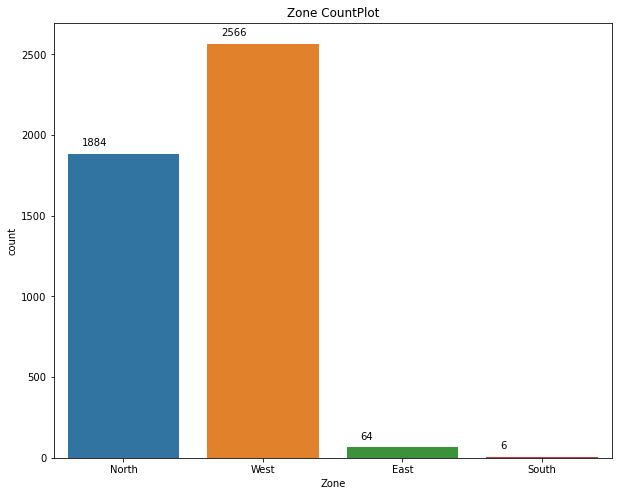
**Marital Status**

Married 0.50 Single 0.28 Divorced 0.18 Unmarried 0.04

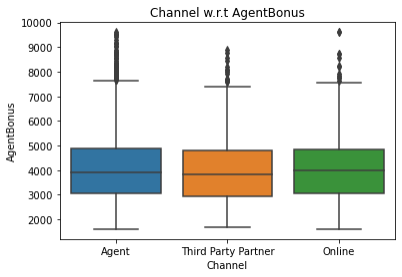
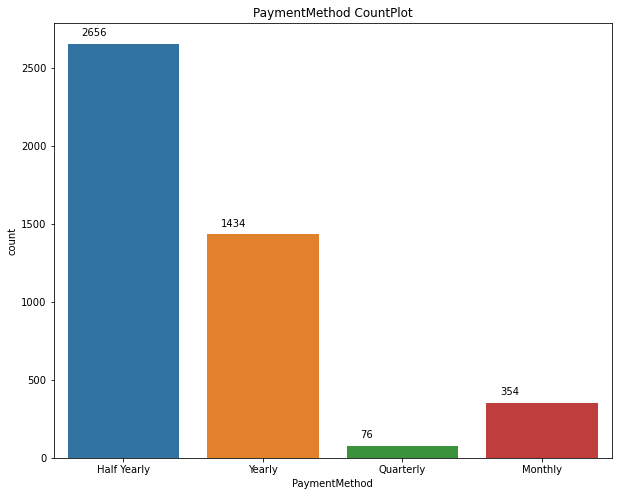
**Around 50% of the customers Are married.**

Figure 2(f) Count Plot -Marital Status

**15**



**Zone**



West 0.57 North 0.42 East 0.01 South 0.00

West Zone brings the most Customers with 57% weightage. Here freelancers have a minute weightage.

Figure 2(g) Count Plot - Zone

**PaymentMethod**

Half Yearly 0.59 Yearly 0.32 Monthly 0.08 Quarterly 0.02

Around 59% of Customers went For half-yearly payment plan

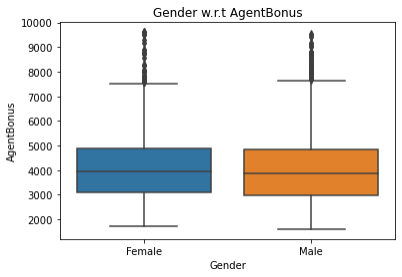
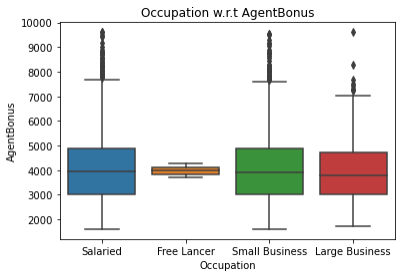
Figure 2(h) Count Plot - PaymentMethod

**Categorical Variables Bivariate Analysis w.r.t Agent Bonus**

· Agent Bonus has a lot of outlier values for every channel with almost similar mean values for all 3 channels.

Figure 3(a) Boxplot – Channel w.r.t AgentBonus

**16**



Almost similar mean value for all Occupations.

NO outliers present for Free Lancer

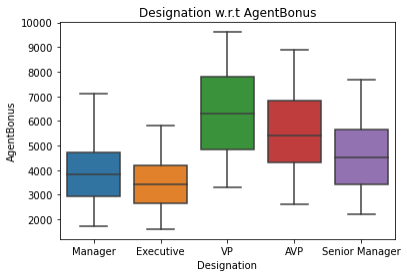
Could be because we have only 2 data points for Free Lancer.

Figure 3(b) Boxplot – Occuapation w.r.t AgentBonus

· Agent Bonus has a lot of outlier values

for both Genders with almost similar mean values for b oth Male and Female.

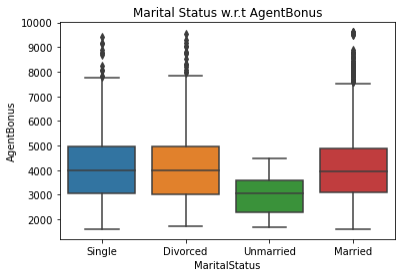
Figure 3(c) Boxplot – Genderl w.r.t AgentBonus

No outliers present.

VP Designation has the highest mean As compared to other Designations.

Figure 3(d) Boxplot – Desgnationl w.r.t AgentBonus

**17**



· Agent Bonus has a lot of outlier values for all MaritalStatus except Unmarried customers.

· With almost similar mean values for all 3 customers except unmarried.

Figure 3(e) Boxplot – MaritalStatus w.r.t AgentBonus

Outliers present only for North and West Zones. Both having almost

Similar means.

No outliers present in East and South Zones possibly due to less Customer traffic from those Zones.

Figure 3(f) Boxplot – Zone w.r.t AgentBonus

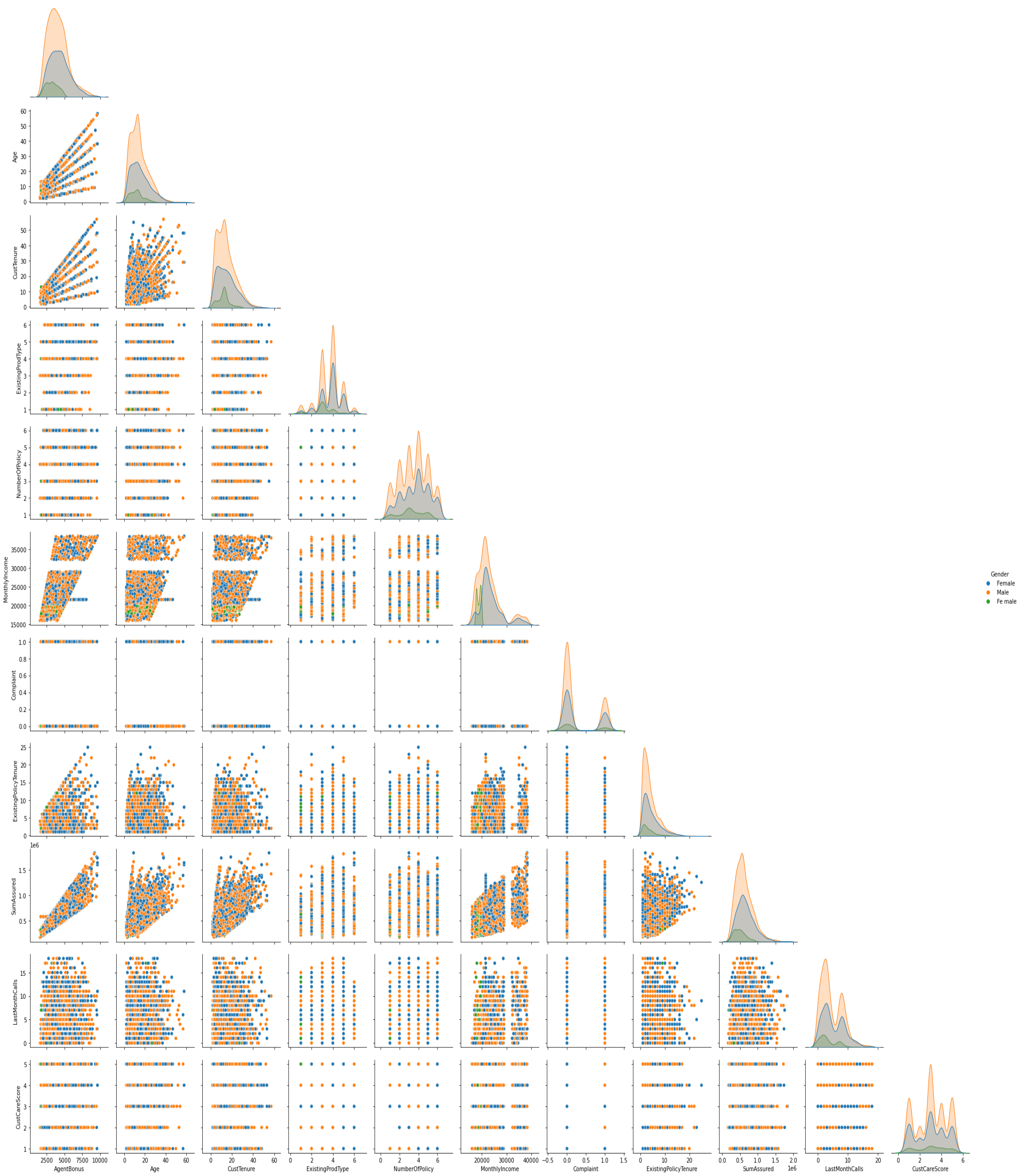
· Outliers present for all

Payment methods chosen by the customer. · Quarterly paying customers having the low

est mean.

Figure 3(g) Boxplot – Channel w.r.t AgentBonus

**18** **Pairplot**



*A pair plot plots the relationships between all numeric variables in a dataset. The diagonal below is the histogram for each variable and shows the distribution. From the below plot, we can observe if there are relationships between every two pair of variables.*

Figure 4 – Pairwise Distribution Plot

**19** **Correlation Heatmap.**



*The correlation coefficient shown in the table below shows the degree of correlation between the two variables represented in X axis and Y axis. It varies between -1 (maximum negative correlation) to +1 (maximum positive correlation).*

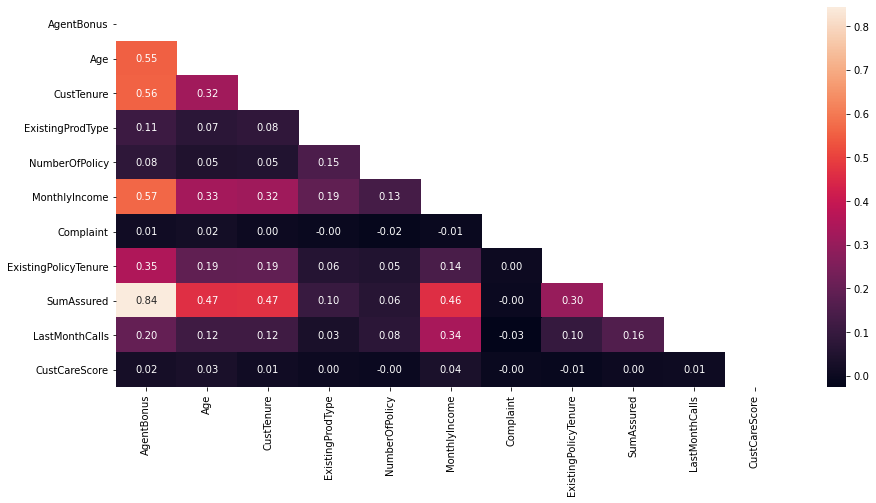


Figure 5 - Correlation Heatmap

· We can observe that there is almost no multicollinearity in the data.

· Complaint and CustCareScore have almost no correlation with any other parameter, hence dropping these columns will not make a difference.

· AgentBonus and SumAssured have high correlation with each other of 0.84. · Here the lighter colors depict high correlation and darker colors depict low

correlation.

1. Outlier Removal is performed but it does not seem as the correct approach as some variables like

2. SumAssured are allowed to have some outliers however our model will be affected if outliers are not removed.

3. We can add new variables like Premium but adding new variables can affect the model, hence not recommended.

4. With this we’ve completed the EDA in the coming exercises we’ll build the model as this is a Classific ation problem, Regression Techniques for model building will be our go-to approach.

5. The data is highly unbalanced eg: Zone, South has less weightage similar for Occupation- Freelancer, more data is needed or upscale the data.

**Thank You!**