**Supplementary data analysis for:**

**Risk Prediction in Life Insurance Industry using Supervised Learning Algorithms**



# 1 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) will involve univariate analyses showing the distributions of the attributes in the data set together with bivariate analyses to gain insights about the relationships which exist between the response variable and certain features. For this purpose, the train data set will be used to build the visualisations.

## Univariate Analysis

### Categorical Variables

Fig. 1 shows bar graphs for the attributes Product\_Info\_1, Product\_Info\_2, Product\_Info\_3, Product\_Info\_5, Product\_Info\_6 and Product\_Info\_7. Product\_Info\_1 shows a higher count for ‘1’. Product\_Info\_2 shows highest count of product ‘D3’ and lowest of ‘B1’. Info ‘26’ accounts for majority of Product\_Info\_3. More than 50% of Product\_Info\_5 exhibits ‘2’. Product\_Info\_6, on the other hand, shows a higher count for ‘3’. Product\_Info\_7 shows the largest proportion for product ‘1’.

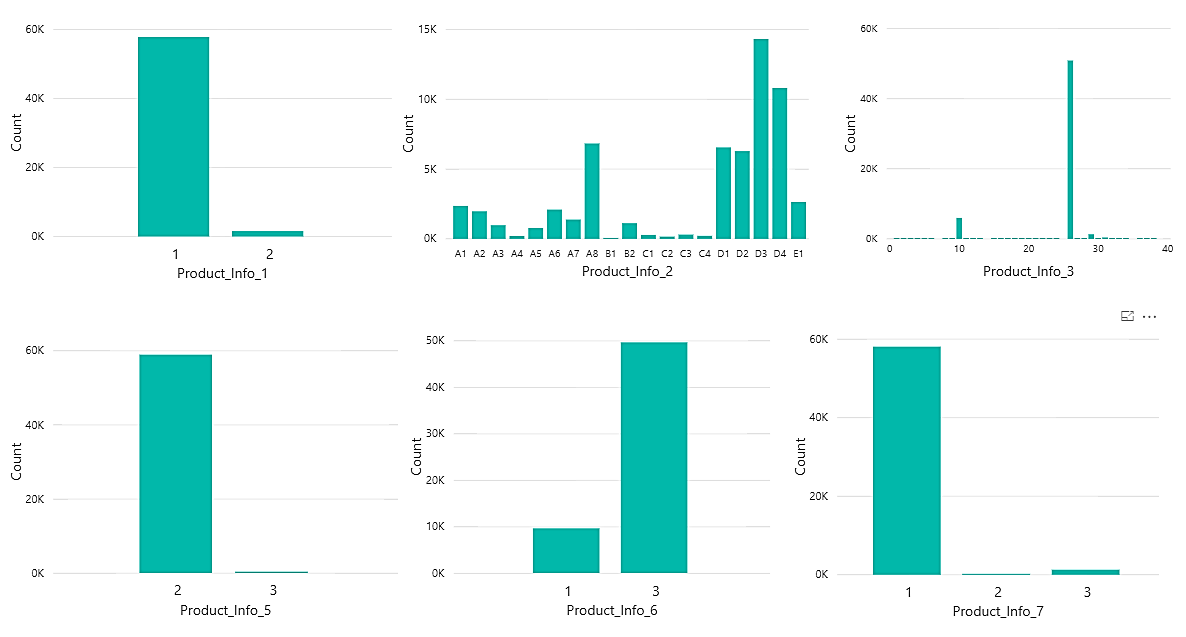


Fig. 1: Product Information 1-3 and 5-6

Fig. 2 shows bar charts for the variables Employment\_Info\_2, Employment\_Info\_3, Employment\_Info\_5 and Family\_Hist\_1. Employment information ‘9’ exhibits the highest frequency for the attribute Employment\_Info\_2 while for Employment\_Info\_3, information ‘1’ shows the highest count and for Employment\_Info\_5, information ‘2’ shows the most occurrence. For the attribute Family\_Hist\_1, data value coded ‘3’shows highest frequency followed by ‘2’and ‘1’.

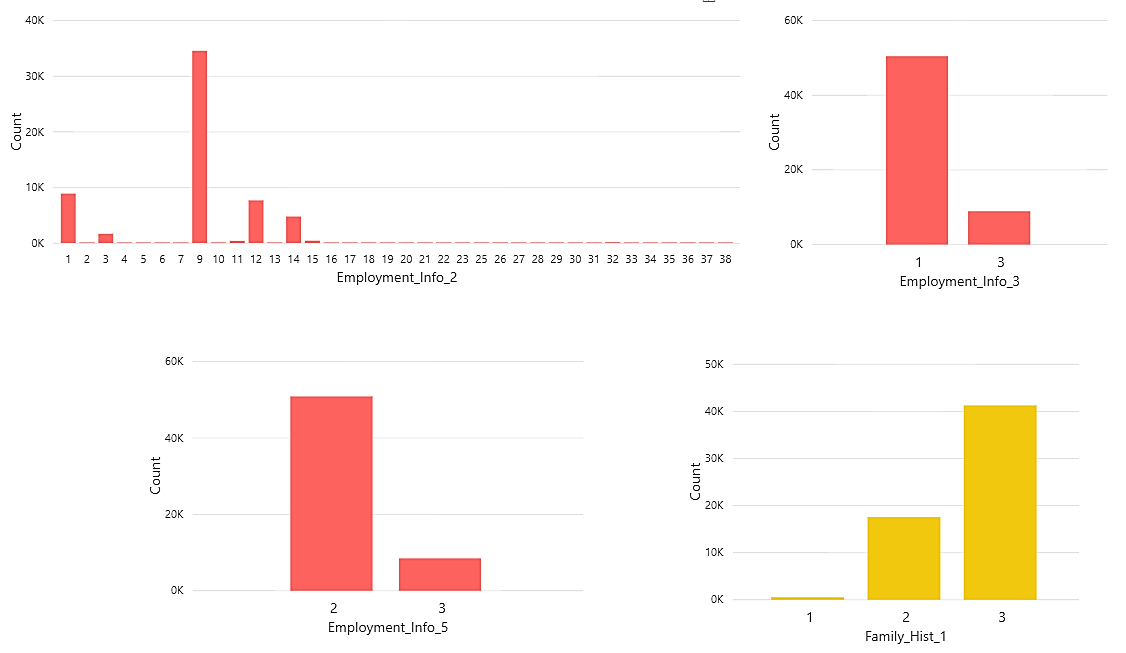


Fig. 2: Employment Information 2, 3, 5 and Family History 1

Fig. 3 shows bar charts for the features InsuredInfo\_1, InsuredInfo\_2, InsuredInfo\_3, InsuredInfo\_4, InsuredInfo\_5, InsuredInfo\_6 and InsuredInfo\_7. Information ‘1’ is most frequent in variables InsuredInfo\_1, InsuredInfo\_5, InsuredInfo\_6 and InsuredInfo\_7. InsuredInfo\_2 exhibits highest count for ‘2’ while InsuredInfo\_4 shows highest count for ‘3’. InsuredInfo\_3 shows highest frequency for ‘8’, followed by ‘3’ and ‘6’ and others.

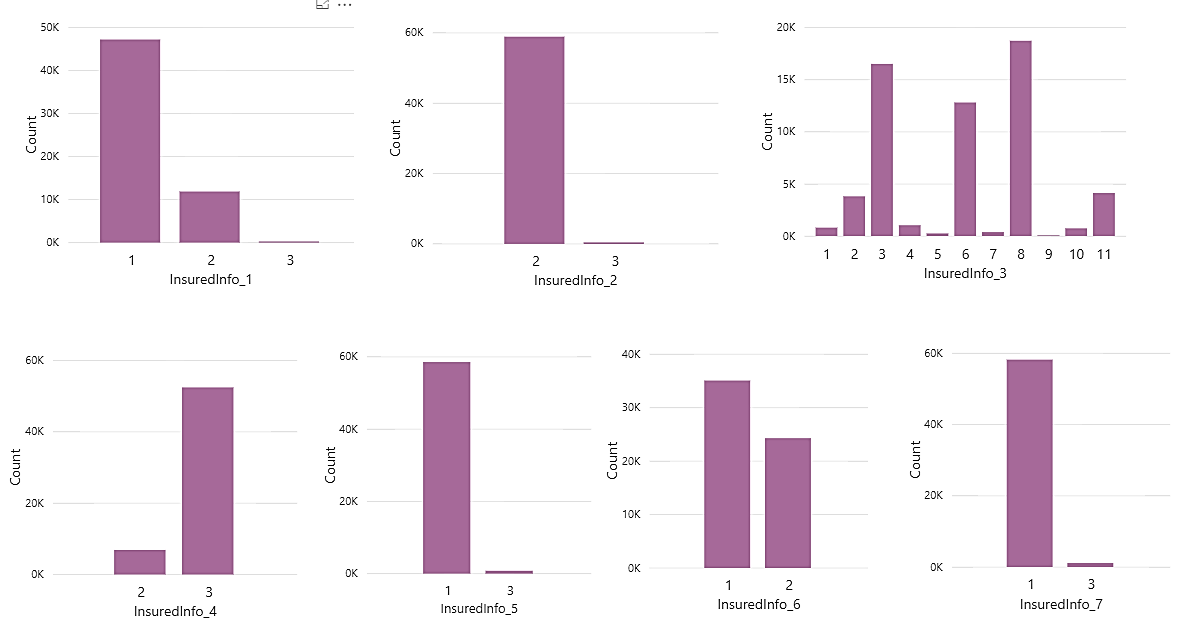


Fig. 3: Insured Information 1-7

Fig. 4 shows bar charts for the attributes Insurance\_History\_1, Insurance\_History\_2, Insurance\_History\_3, Insurance\_History\_4, Insurance\_History\_7, Insurance\_History\_8 and Insurance\_History\_9. Insurance\_History\_1, Insurance\_History\_8 and Insurance\_History\_9 have higher count for data value coded as ‘2’ while Insurance\_History\_2, Insurance\_History\_4 and Insurance\_History\_7 have higher count for data value coded as ‘1’. On the contrary, coded value ‘3’ shows highest count for Insurance\_History\_3.

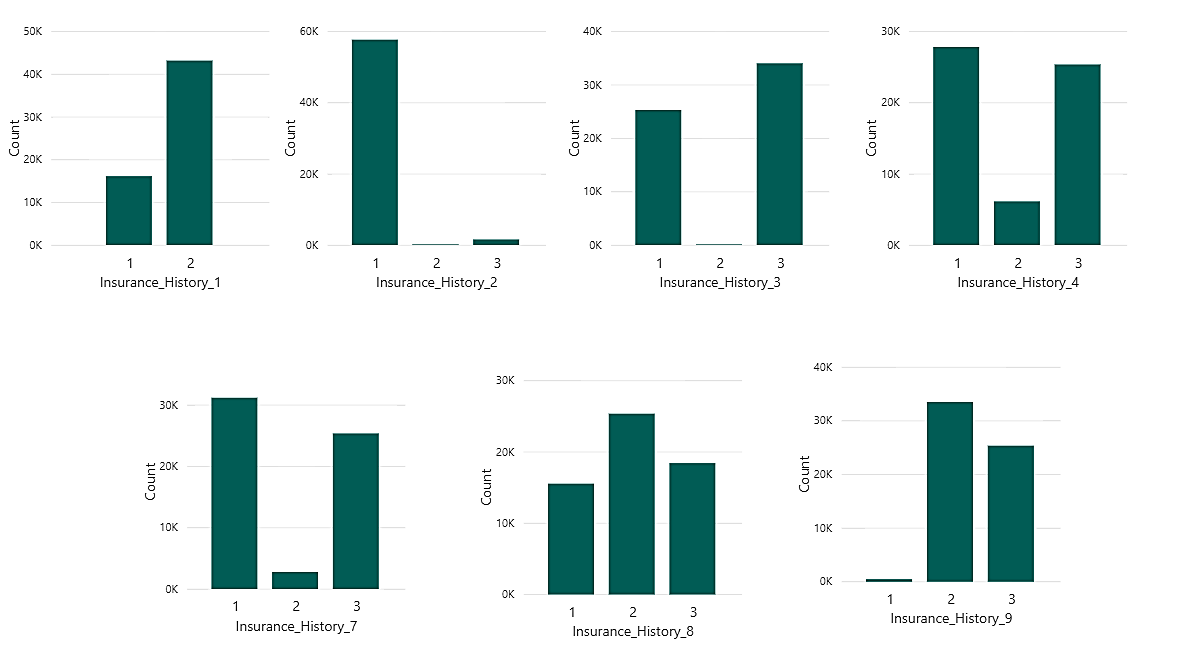


Fig. 4: Insurance History 1-4 and 7-9

Fig. 5 illustrates the distribution for Medical\_History features. Medical\_History\_2 shows highest frequency for value coded ‘112’. Medical\_History\_3, Medical\_History\_4, Medical\_History\_7, Medical\_History\_8 and Medical\_History\_9 have highest occurrence of ‘2’. Medical\_History\_5 shows highest count for ‘1’ while Medical\_History\_6 shows highest frequency for value ‘3’.

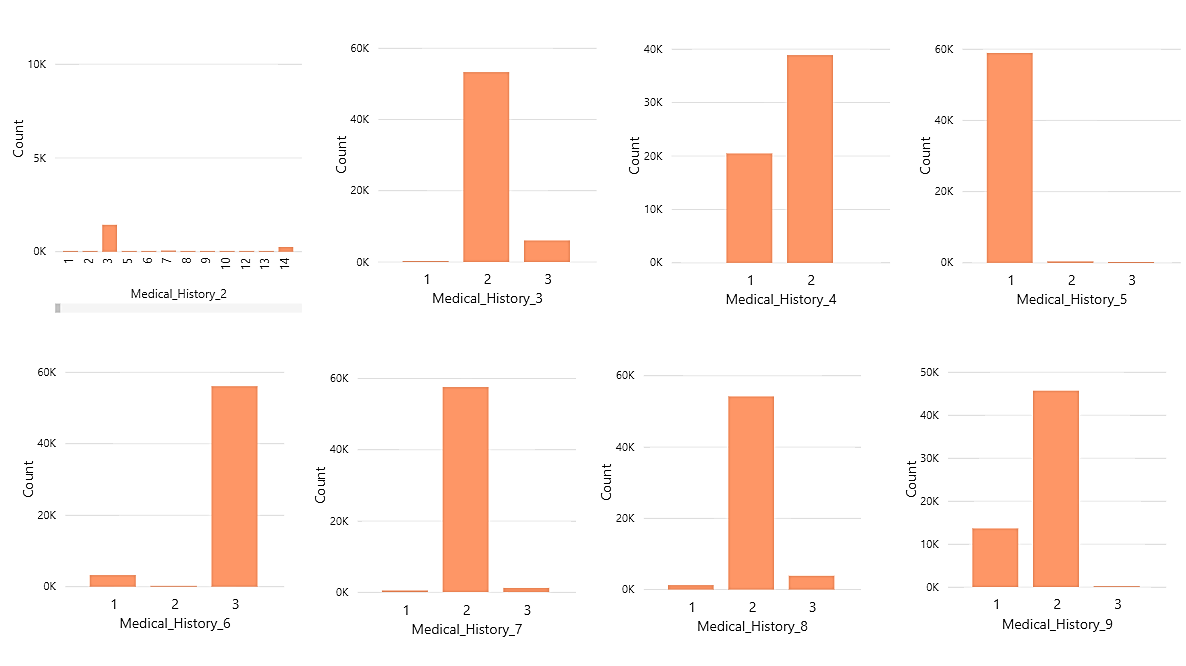


Fig. 5: Medical History 2-9

Medical\_History\_11, Medical\_History\_13, Medical\_History\_14 and Medical\_History\_17 show high frequency for coded value ‘3’ while Medical\_History\_16, Medical\_History\_18 and Medical\_History\_19 have higher occurrences for ‘1’. Besides, Medical\_History\_12 has higher count for coded value ‘2’ as depicted by Fig. 6.

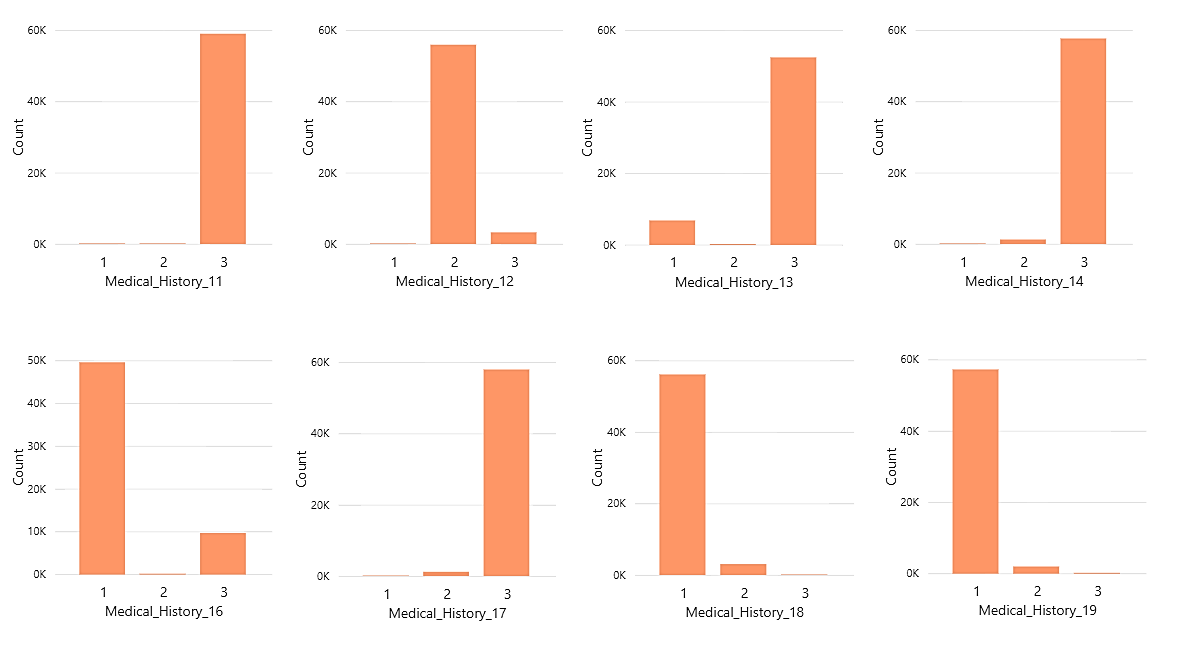


Fig. 6: Medical History 11-14 and 16-19

Medical\_History\_20 and Medical\_History\_22 depict highest counts for ‘2’ as illustrated by Fig. 7. Medical\_History\_21, Medical\_History\_25 and Medical\_History\_28 show higher counts for ‘1’ while Medical\_History\_23, Medical\_History\_26 and Medical\_History\_27 have highest frequencies for ‘3’.

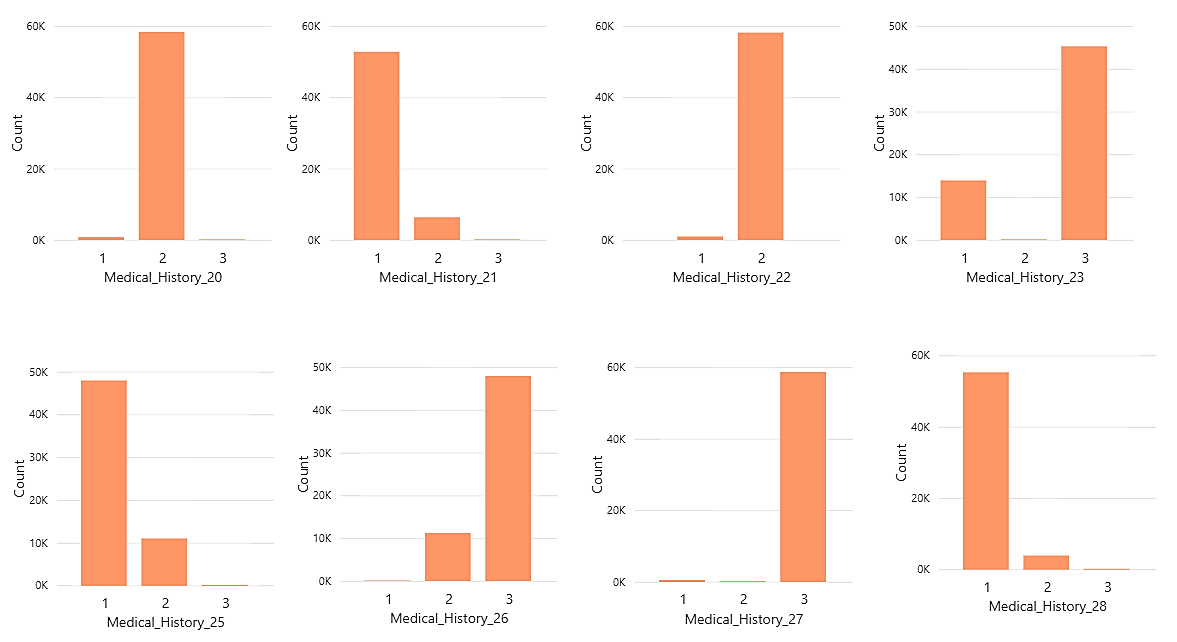


Fig. 7: Medical History 20-23 and 25-28

Fig. 8 illustrates the distributions for Medical\_History 29 to 31 and 33 to 37. The feature Medical\_History\_35 shows a higher count for coded value ‘1’. Medical\_History\_30, Medical\_History\_36 and Medical\_History\_37 show higher occurrence for coded value ‘2’. There is a higher count of ‘3’ as exhibited by the variables Medical\_History\_29, Medical\_History\_31, Medical\_History\_33 and Medical\_History\_34.

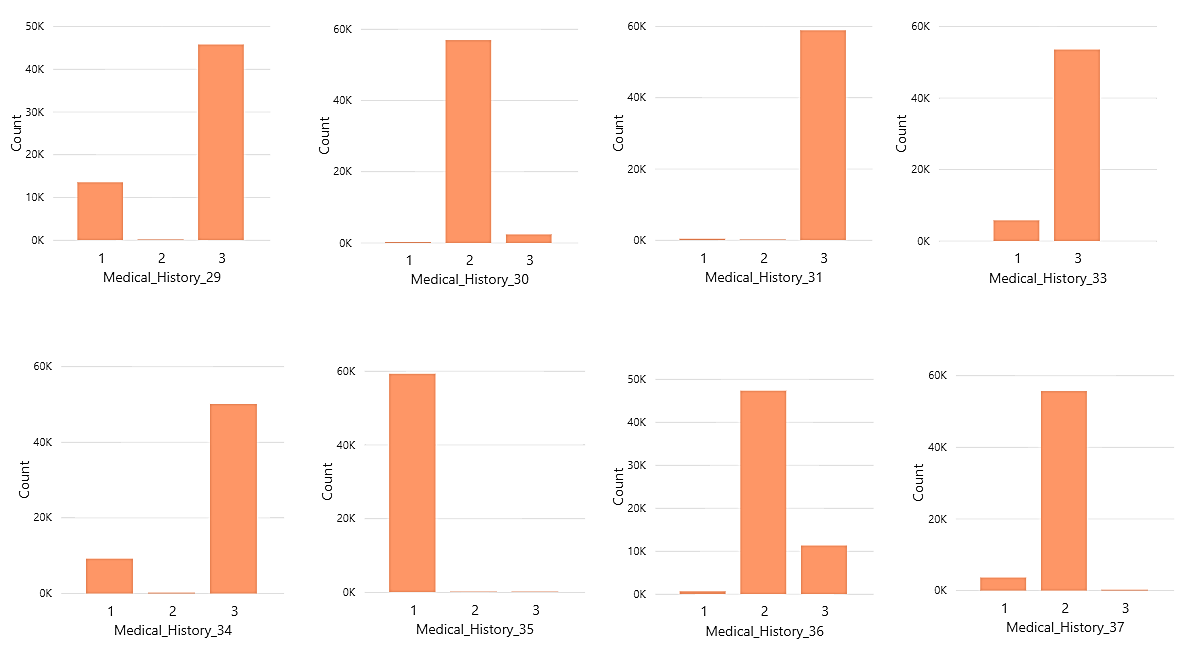


Fig. 8: Medical History 29-31 and 33-37

Fig. 9 illustrates that the features Medical\_History\_38 and Medical\_History\_41 have highest count for the value ‘1’. Medical\_History\_39 and Medical\_History\_40, on the contrary, show higher counts for the coded value ‘3’.

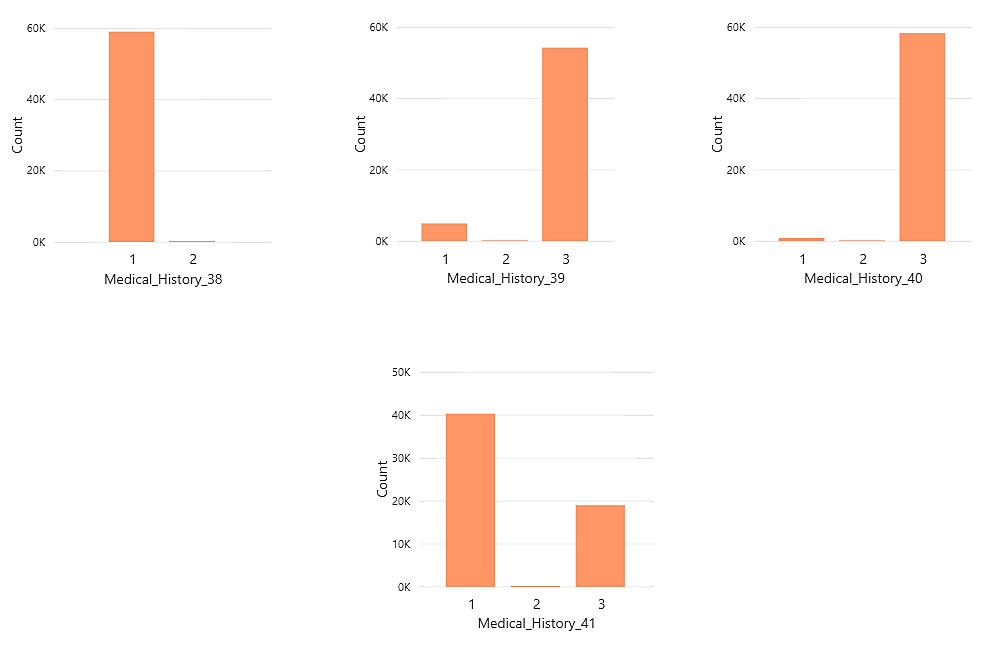


Fig. 9: Medical History 38-41

### Continuous Variables

The figures that follow gives a representation of the continuous attributes present in the data set. Fig. 10 shows area charts for the features Ins\_Age, Ht, Wt and BMI. The distribution of each attribute can be observed. Ins\_Age and BMI shows an approximate normal distribution while Ht and Wt has a slight skewed distribution.

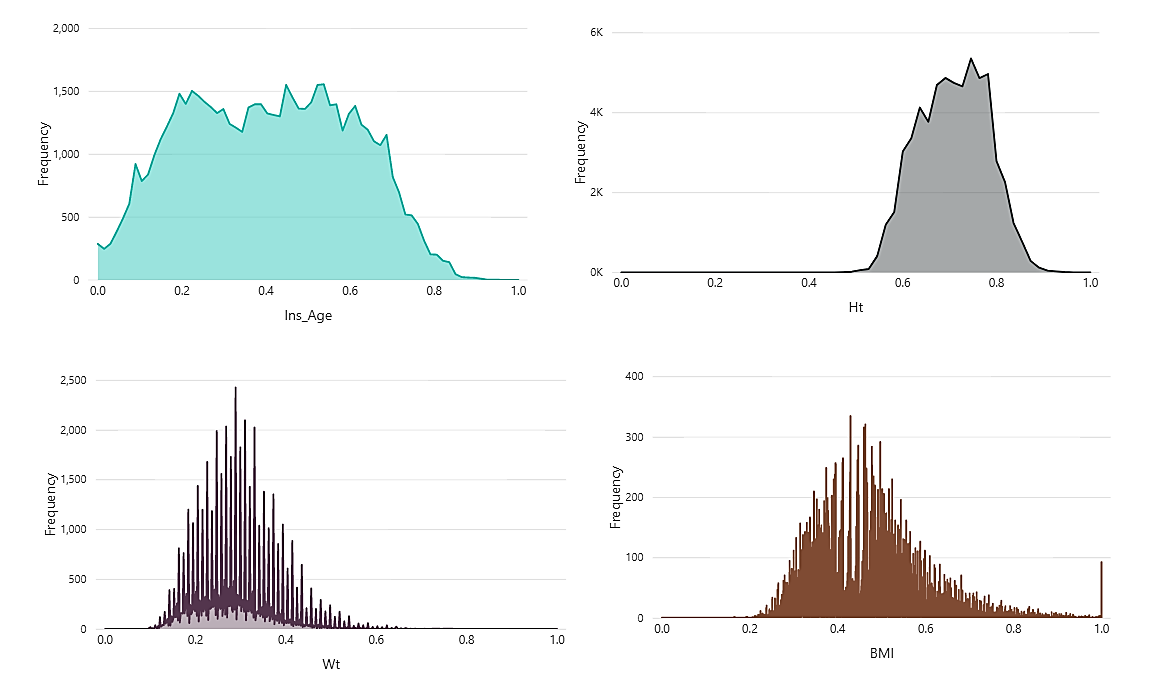


Fig. 10: Age, Height, Weight and BMI distributions

Fig. 11 illustrates the area charts for employment information 1,4 and 6. All three attributes depicted have skewed distributions showing peaks at certain specific values throughout the data values.

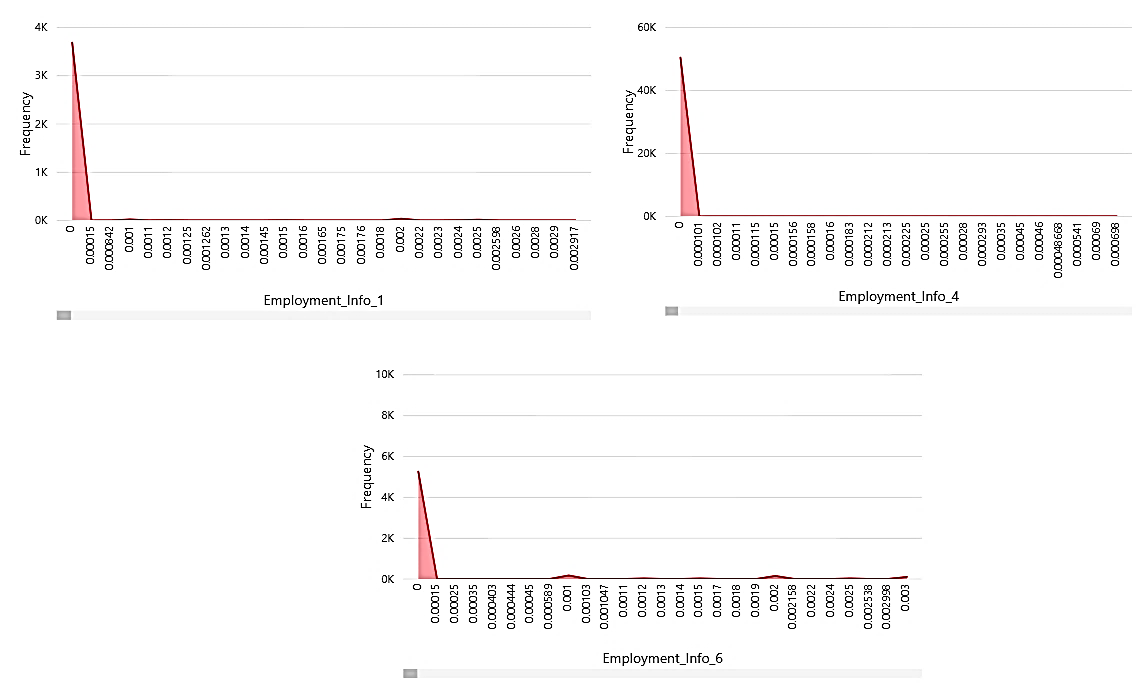


Fig. 11: Employment Information 1, 4 and 6

Product\_Info\_4 and Medical\_History\_1 show left skewed distributions as depicted by the line graphs in Fig. 12.

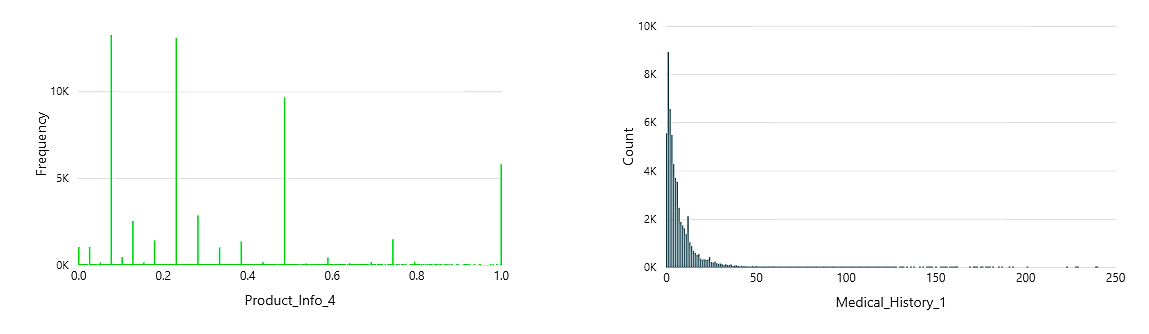


Fig. 12: Product Information 4 and Medical History 1

### Response Variable

The horizontal bar chart in Fig. 13 illustrates the frequency distribution of the response variable in the data set. This is an ordinal measure of risk varying from 1 to 8, with 1 showing the lowest risk level and 8 showing the highest risk level. Out of 59381 instances in the train data set, 19489 cases have risk level 8, which is the highest occurring frequency. This is followed by risk level 6, with 11233 cases reported. The lowest occurring risk level belongs to level 3, with 1013 instances.

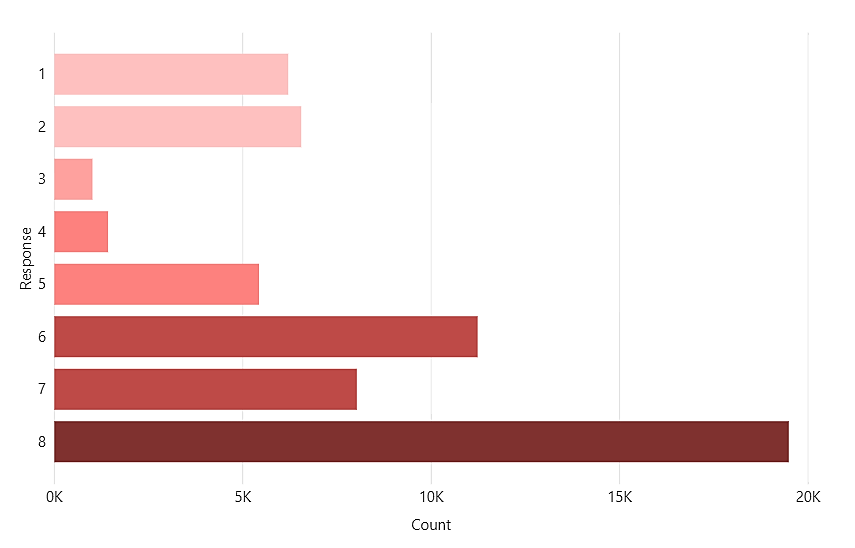


Fig. 13: Response Variable

## Bivariate Analysis

The distributions of the demographic variables in the data set with the response variable is analysed by using line charts and box plot. Four hypotheses are outlined regarding the data set, which are explained below.

*Hypothesis 1: Does age have a significant impact on the risk level?*

The attribute Ins\_Age shows a significant difference between risk levels 1 to 8. Response groups 3,4 and 5 record the lowest counts due to them having smaller sizes in the data set, refer Fig. 14.

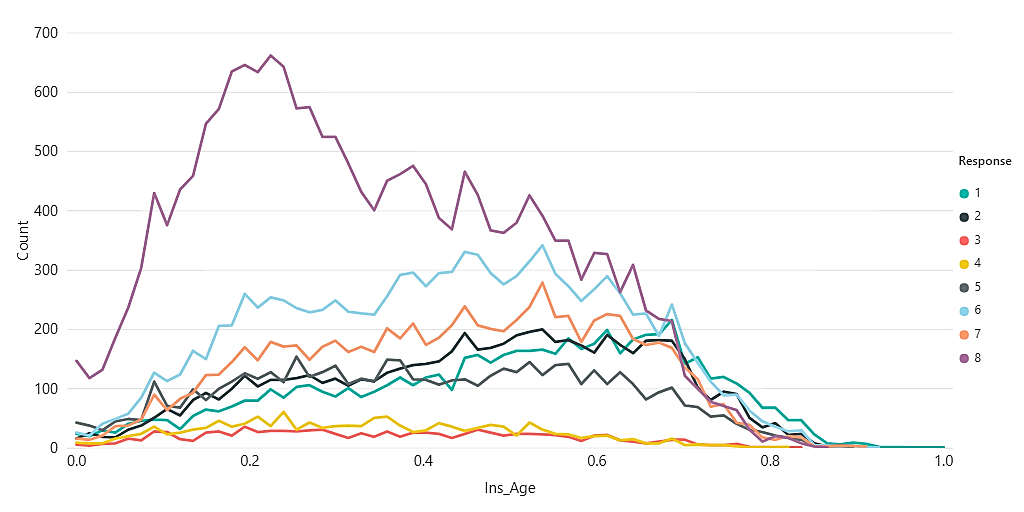


Fig. 14: Age by Response

*Hypothesis 2: Does height vary according to the different response groups?*

The attribute Ht also shows significant differences among the different response groups. Detailed counts from Fig. 15 below shows that for height value 0.78, response level 6 shows a higher count compared to response level 8, which has higher counts for almost all height values.

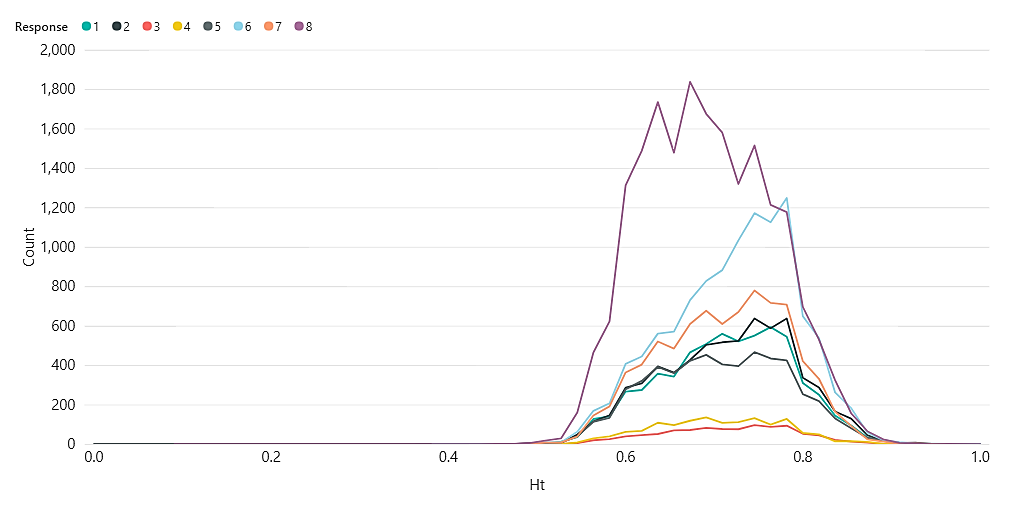
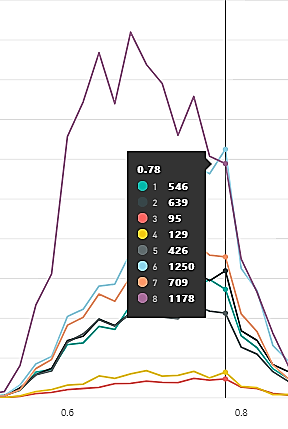


Fig. 15: Height by Response

*Hypothesis 3: Does weight have an influence on the risk level?*

The weight distribution for all response groups lies between 0 to 0.8. However, response group 8 shows no count beyond weight 0.5 while response group 1 still shows values beyond 0.5, refer Fig. 16.

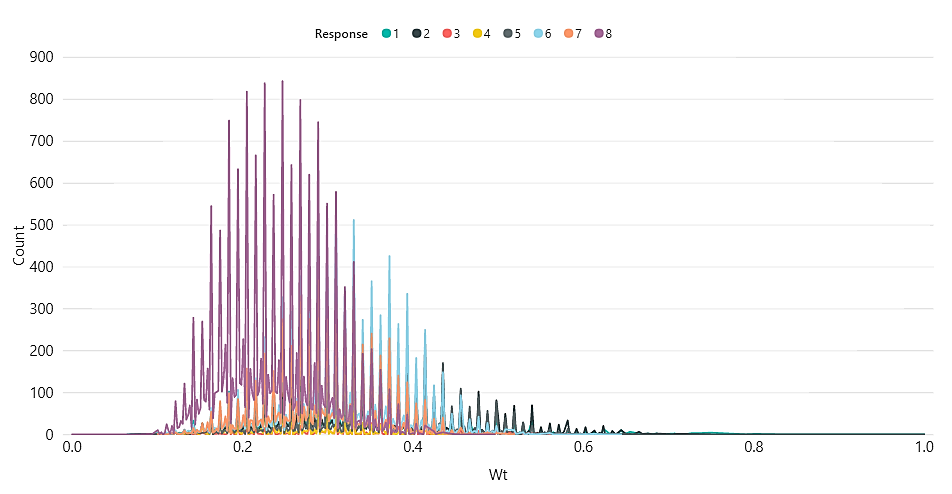


Fig. 16: Weight by Response

*Hypothesis 4: What is the BMI distribution across the different response groups?*

The box plot below shows (Fig. 17) the distribution of BMI for the different responses. Response 8 and 4 have lower average BMI values as compared to the other response groups. Response 2 and 5 have highest average BMI among all the response groups.

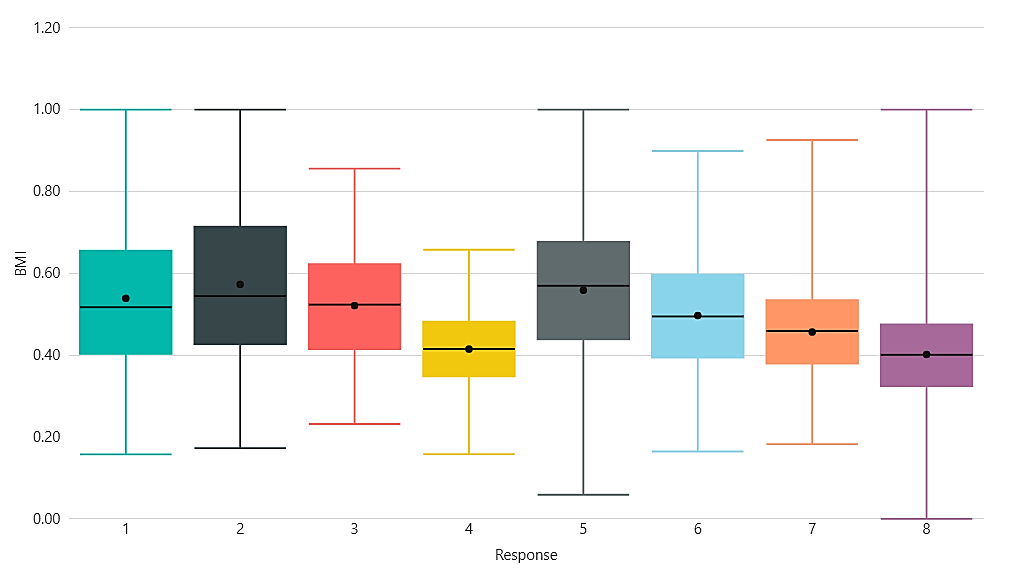


Fig. 17: BMI by Response