AUTO INSURANCE CLAIMS

EXPLORATORY DATA ANALYSIS

PUSHKAR CHOUGULE

STATISTICAL QUESTION/HYPOTHESIS

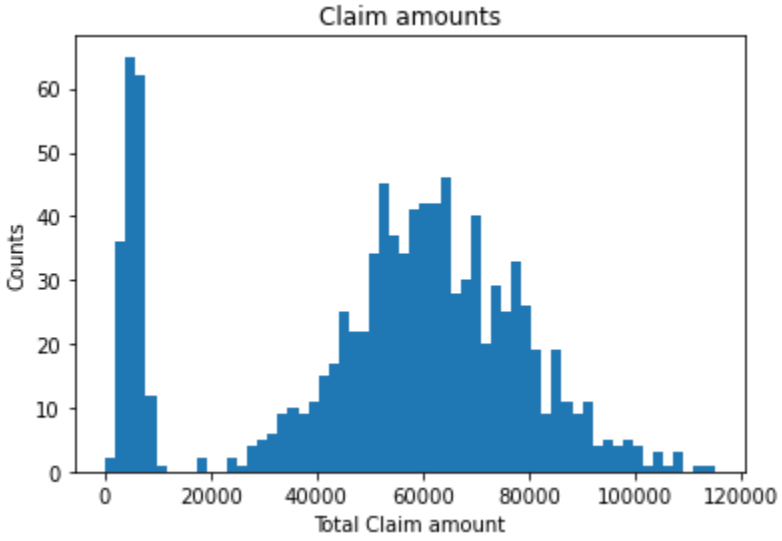
- What variables are more correlated with accident incidents?/ What Factors are statistically significant amongst the drivers involved in accidents?
- Hypothesis: There is an effect of the multiple predictor variables (listed in next slide) on the outcome as drivers being susceptible to accidents
- Null Hypothesis: The predictor variables do not indicate whether drivers are potentially risky.

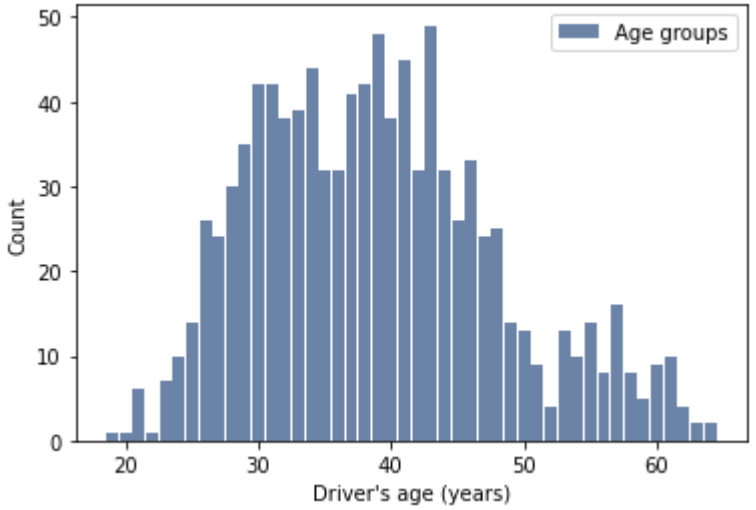
Based on this assumption, compute the probability of the apparent effect. P-Value. If the P-value is low for selected predictor variables, the null hypothesis is unlikely to be true.

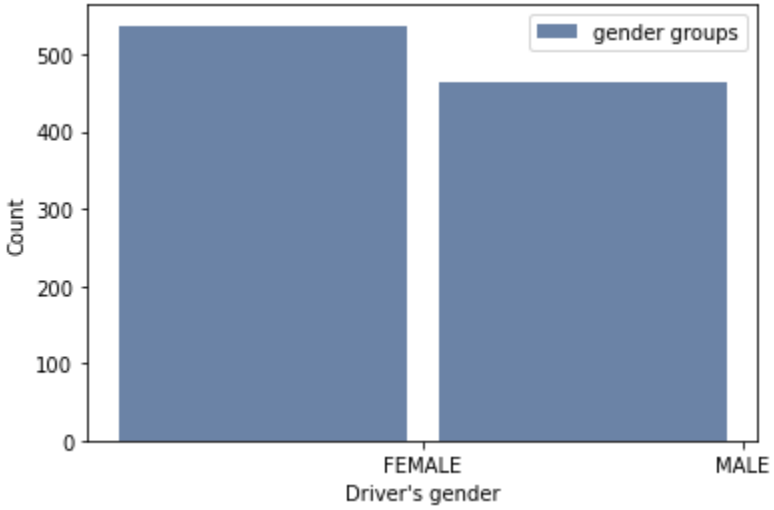
VARIABLES

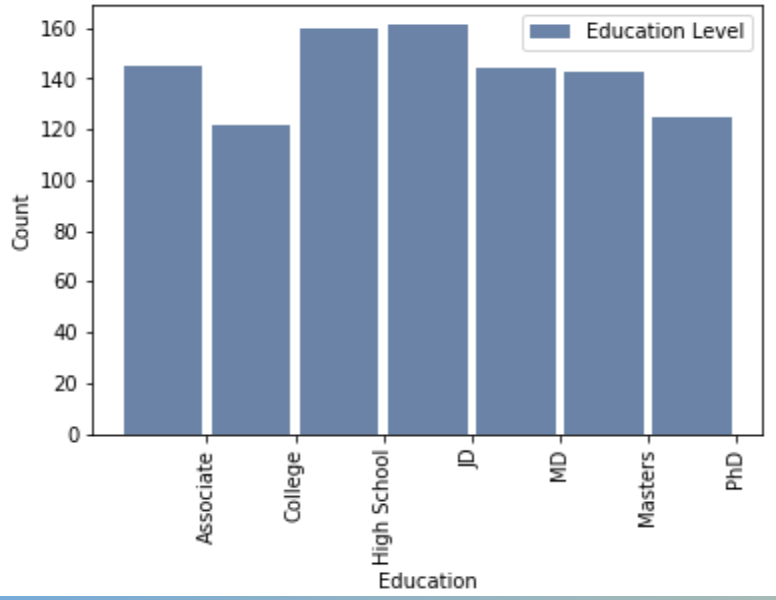
- The datasets consists of several variables, out of which below have been selected as predictor variables and one target variable as total claim amounts:
- Brief description of Variables selected for Analysis. Variable names are fairly self descriptive.
- 1) age : Age of the Insured
- 2) policy_state : Policy state / Insured's home state
- 3) insured_sex : Insured's gender
- 4) insured_education_level : Insured's education
- 5) insured_occupation : Insured's occupation
- 6) insured_relationship : Insured's relation
- 7) incident_state : Accident incident's state
- 8) total_claim_amount : Total claim amount for the accident incident
- 9) auto_make : Auto Make of the involved vehicle
- 10) auto_model : Auto Model of the involved vehicle

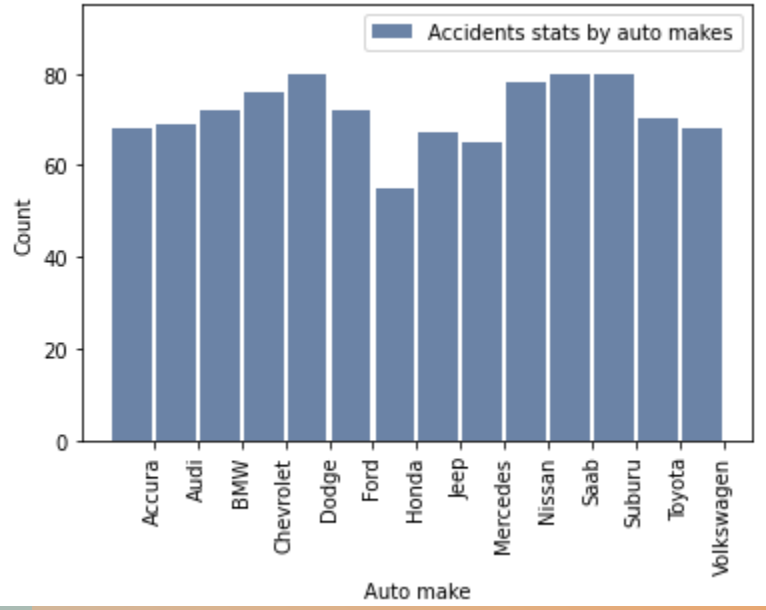
HISTOGRAM OF VARIABLES

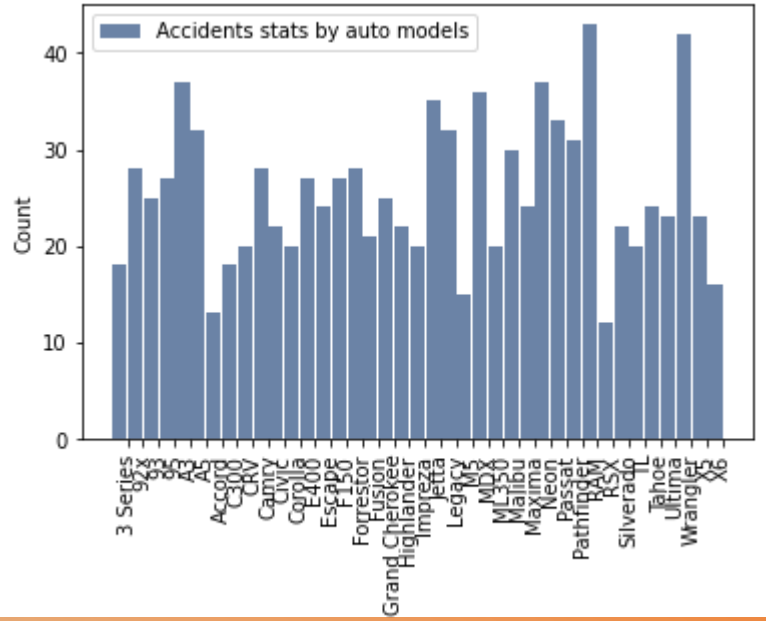




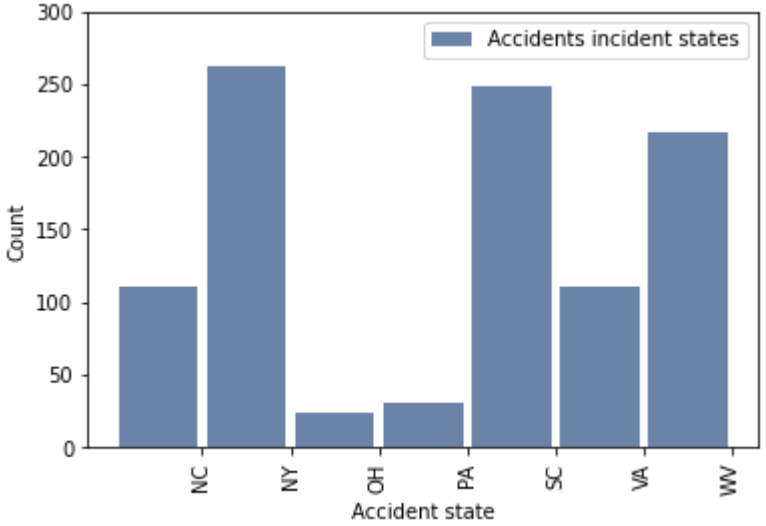


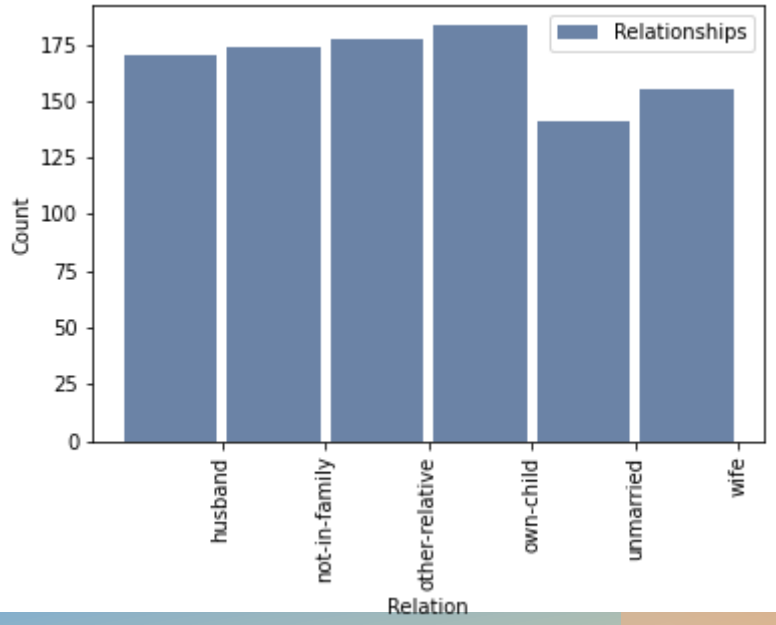


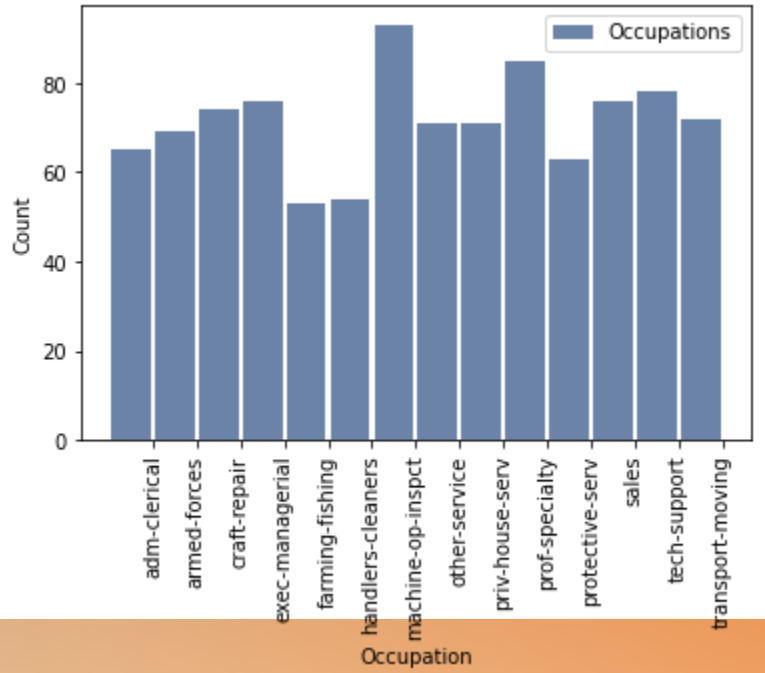




HISTOGRAM OF VARIABLES (CONTINUED..)







DESCRIPTIVE CHARACTERISTICS ABOUT VARIABLES

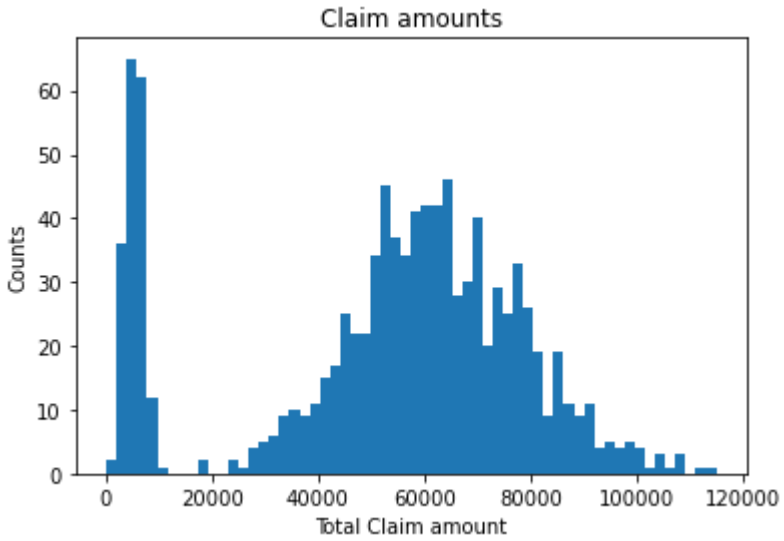
	age	total_claim_amount
count	1000.000000	1000.00000
mean	38.948000	52761.94000
std	9.140287	26401.53319
min	19.000000	100.00000
25%	32.000000	41812.50000
50%	38.000000	58055.00000
75%	44.000000	70592.50000
max	64.000000	114920.00000

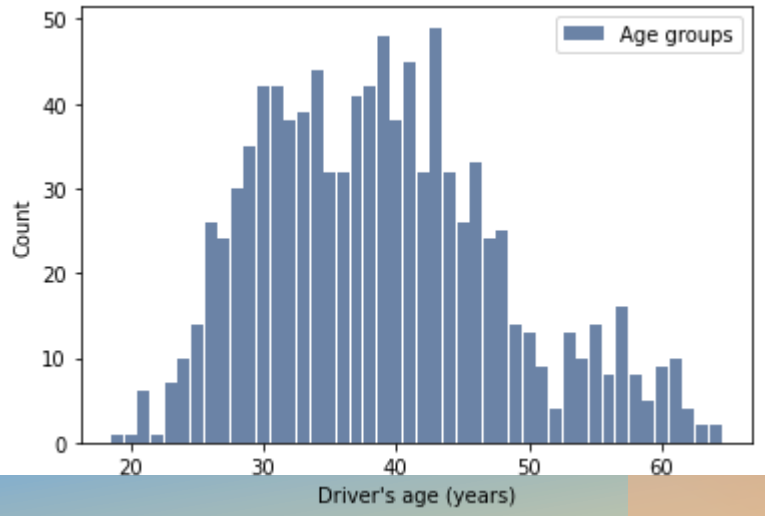
```
mode value for age: 43 years
mode value for total_claim_amount: $ 59400
```

Age and total_claim_amount are the only 2 numeric variables I selected in the vehicle insurance claims analysis. Hence using describe() and mode() methods to display various properties of these variables. These provide respective total counts, mean, spreads like standard deviation, along with 1st quartiles (25%), medians (50%) and 3rd quartiles (75%) as well as min / max values.

For all other selected categorical variables, Histograms and PMFs consist of corresponding values for each column variable. We won't be able to describe mean, mode, median, quartiles, minimum / maximum etc. since categorical variables are unordered.

DESCRIPTIVE CHARACTERISTICS ABOUT VARIABLES AND POSSIBLE VISUAL SKEWS / OUTLIERS



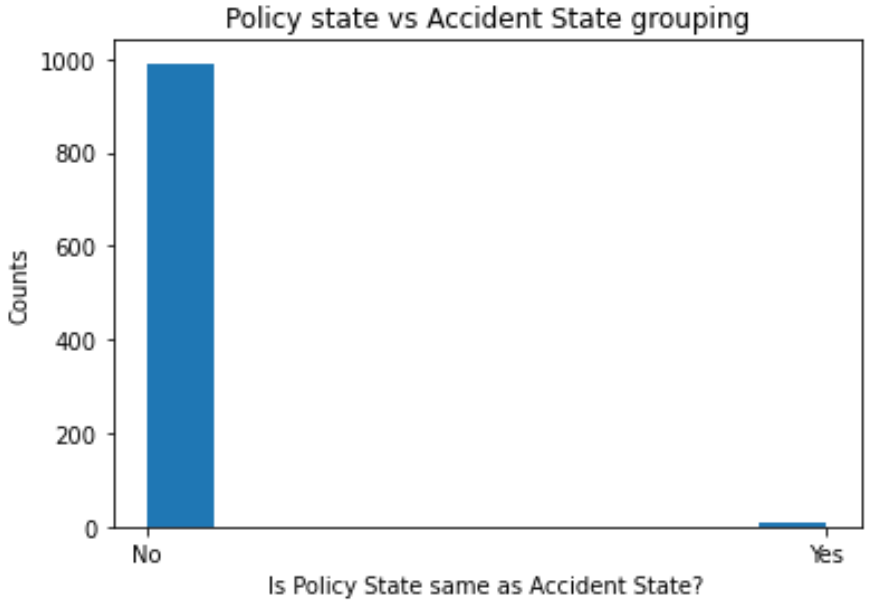


- Total claim amounts distribution shows normal distribution between $20,000 to $120,000 range. But also appears to be somewhat negatively skewed with many records on the lower end of the distribution (i.e. below $20,000)
- Age group distribution appears to be mostly normal with slight bit of Skew on the higher end of the age groups i.e. Slightly Positively Skewed – slightly longer tail towards right.
- Age is understandably skewed to the right as this sample is drivers age 19 and above

OUTLIERS

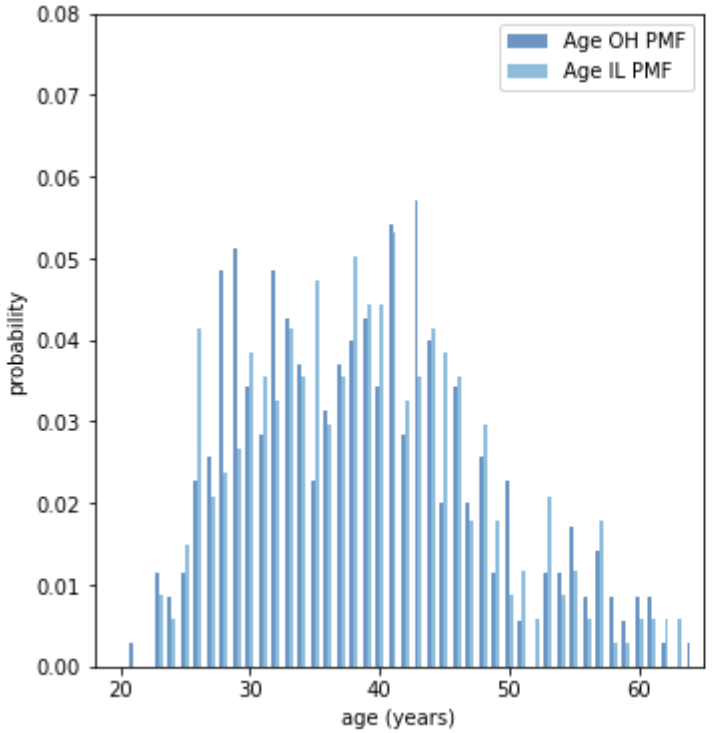
- Looking at the histogram for total claim amounts, it is almost tempting thought to consider the claims with total amounts less than $20,000$ dollars as outliers and drop them and rest of the plot would look like close to Normal distribution. But looking at the total counts of claims below $20,000$ dollars it turns out to be 180 claims, which is 18% of the total available data (1000 records). So, I am holding off on that thought for now.
- Using the IQR range calculations to detect outliers in Total claim amount field
- total_claim_amount_lower = (41812.5 - (70592.5 - 41812.5) * 1.5) = $ -1357.50
- total_claim_amount_upper = (70592.5 + (70592.5 - 41812.5) * 1.5) = $113,762.5
- As we can see, lower end value turns out to be negative amount, which won't be valid for total claim amount. So, effectively all the claim amount records above 0 will need to be considered. For the upper end of range, we end up excluding records with total claim amount greater than $113,762.5

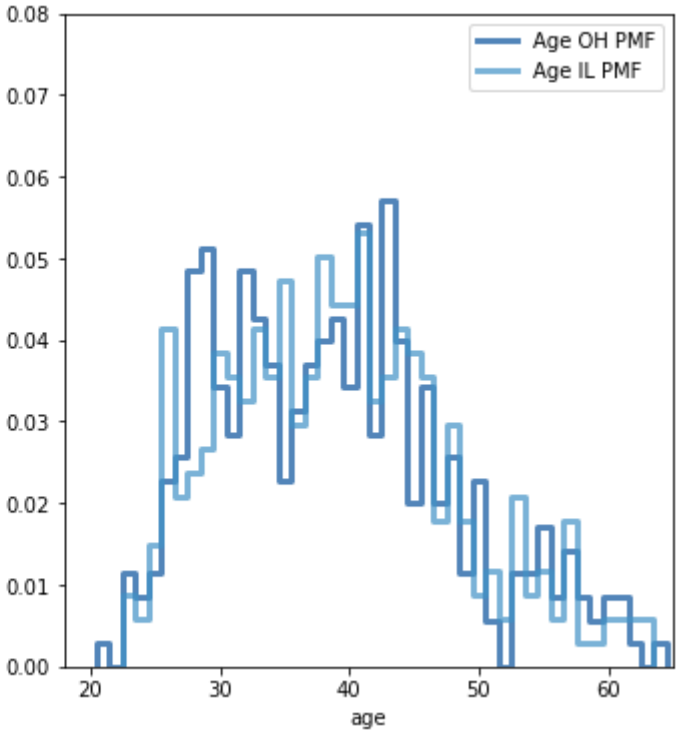
DESCRIPTIVE CHARACTERISTICS ABOUT VARIABLES(CONTINUED..)



- As we see only 9 records out of 1000 records have policy holder's state same as accident incident state.
- This statistics helps confirm the analysis that policy holders driving outside the home state are more accident prone and can be flagged as risky.

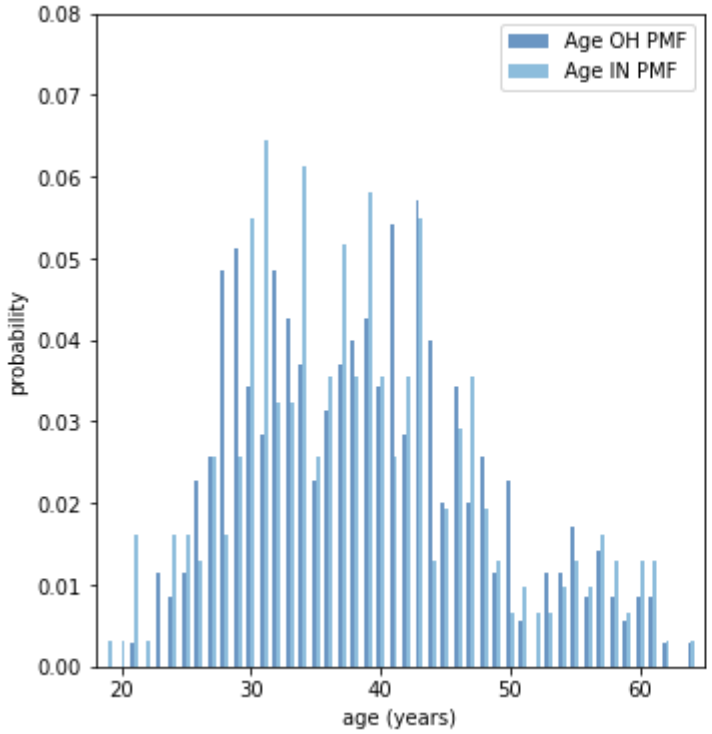
PROBABILITY MASS FUNCTION FOR DRIVERS FROM DIFFERENT STATES – COMPARE BY AGE GROUPS (PART 1)

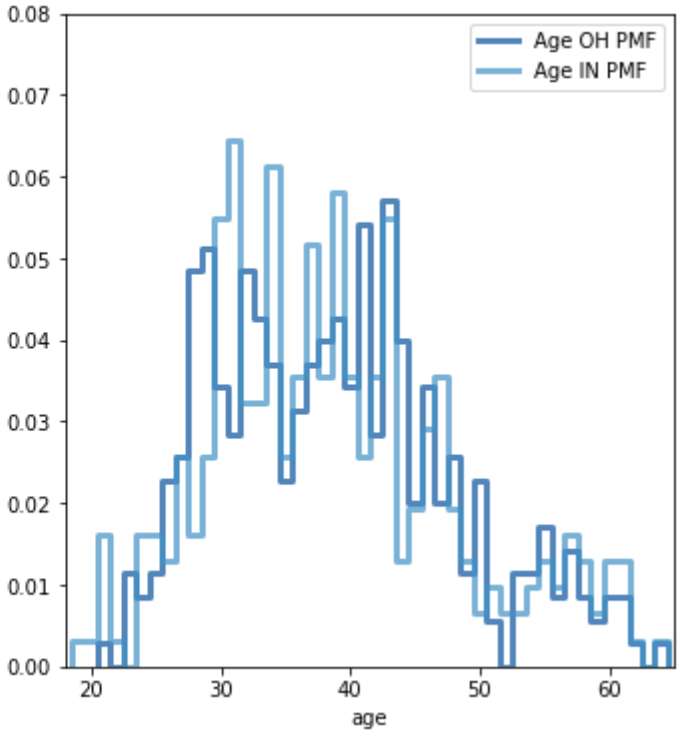




- State of Ohio has a greater number of drivers in the age group of 26 years to 46 years compared to Illinois state drivers involved in accidents

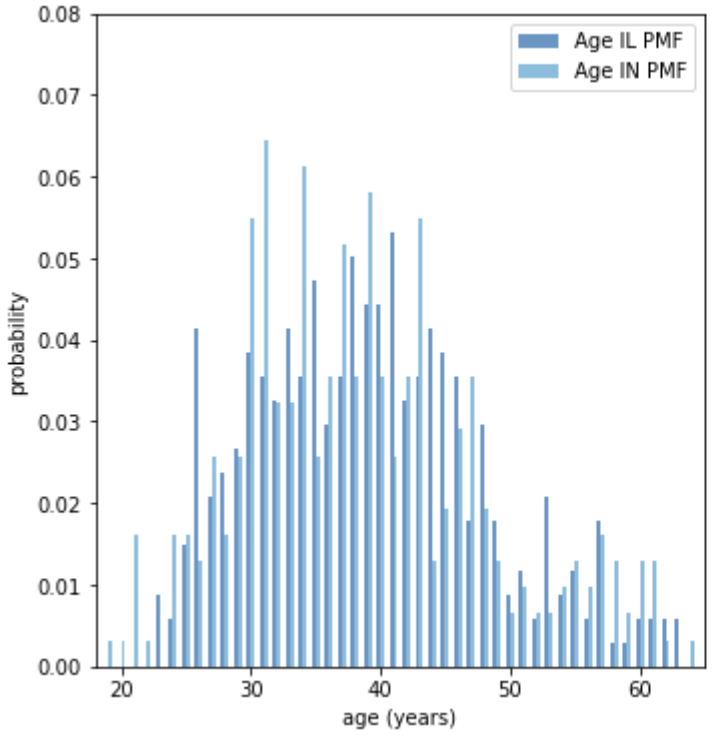
PROBABILITY MASS FUNCTION FOR DRIVERS FROM DIFFERENT STATES – COMPARE BY AGE GROUPS (PART 2)

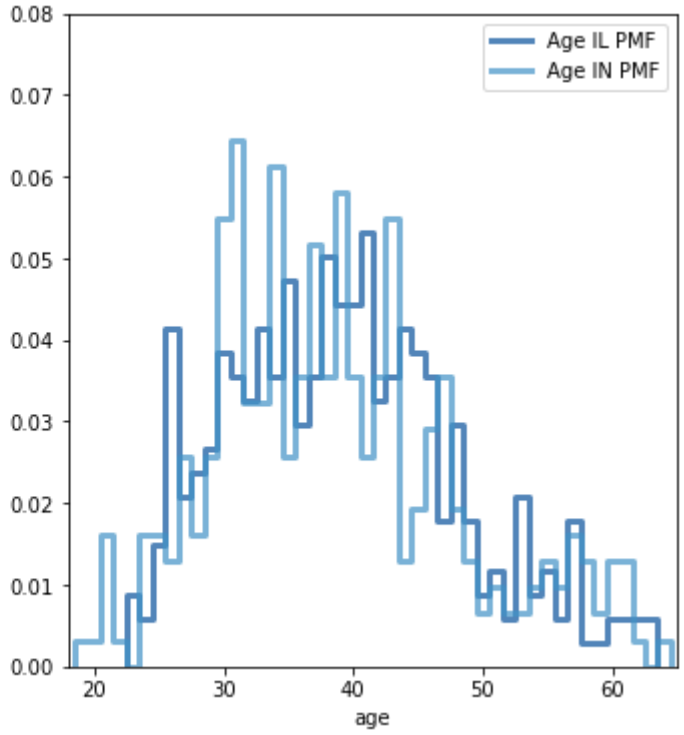




- State of Indiana has a greater number of drivers in the age group of 28 years to 44 years compared to Ohio state drivers involved in accidents

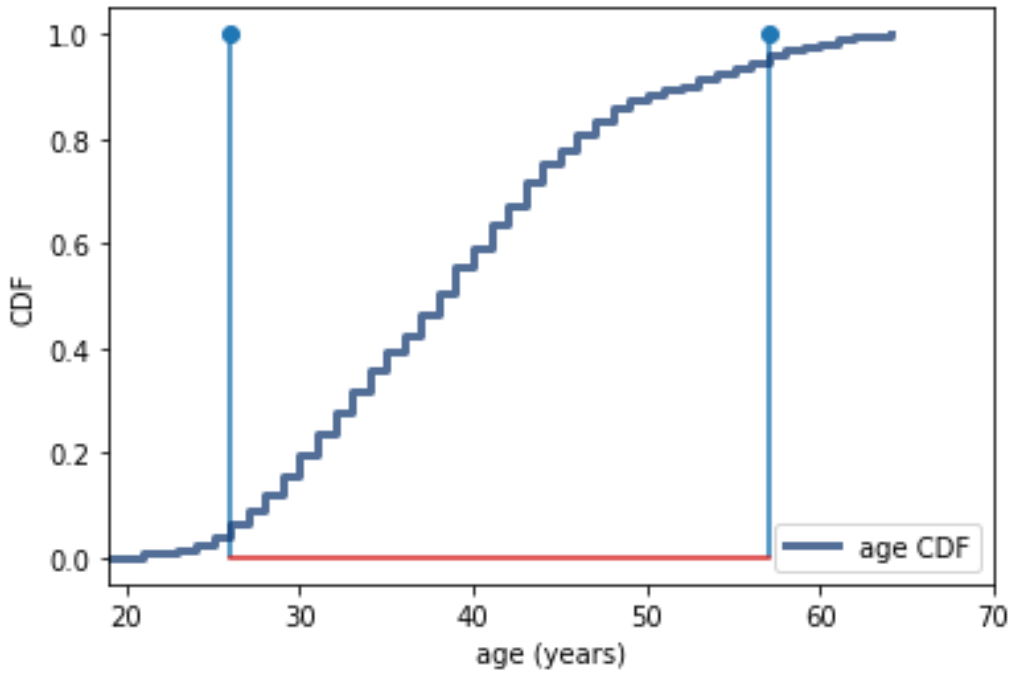
PROBABILITY MASS FUNCTION FOR DRIVERS FROM DIFFERENT STATES – COMPARE BY AGE GROUPS (PART 3)





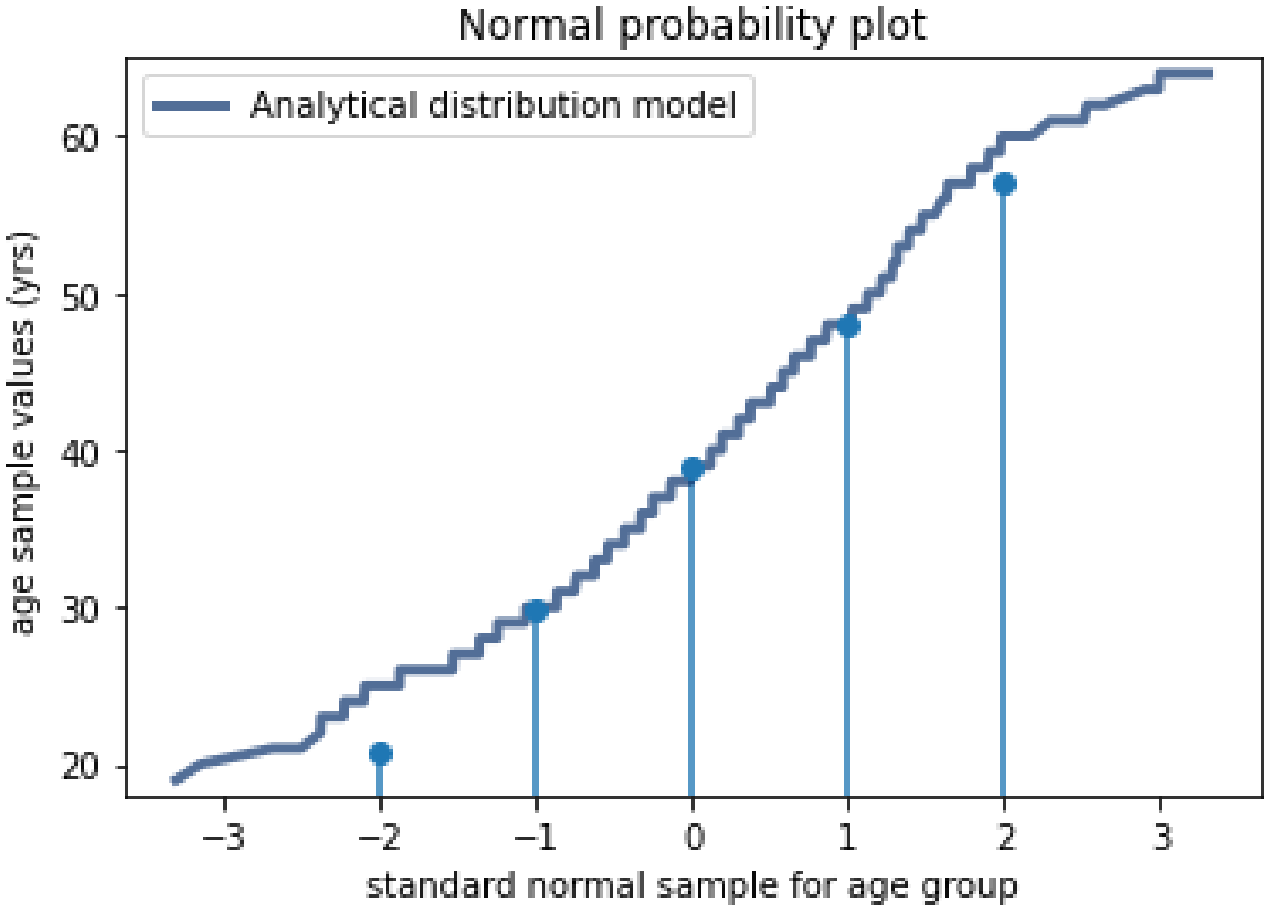
- State of Indiana has a greater number of drivers in the age group of 29 years to 43 years compared to Illinois state drivers involved in accidents

CDF



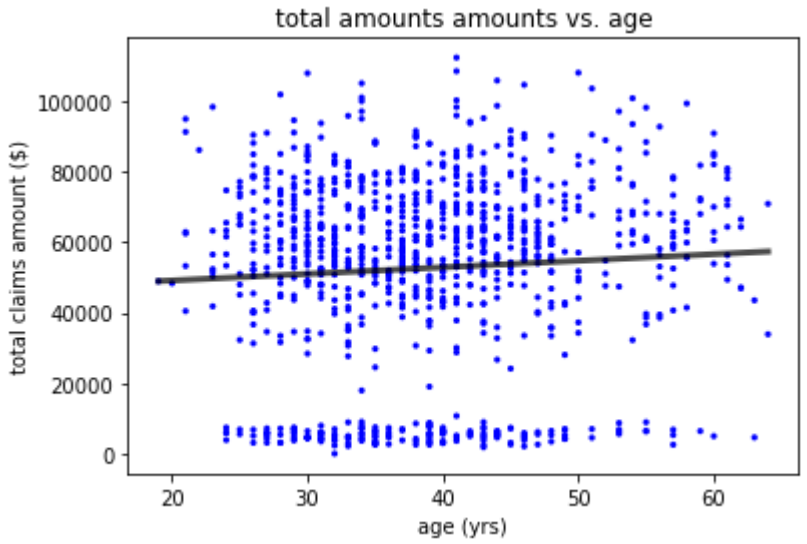
- Overall, This CDF shows majority of the insured drivers involved in the accident fall in the 26 years (5%) to 57 years (95%) age group i.e. 90% confidence interval. So, it would be the risky group to be considered, while underwriting the policy and determining premium rates.
- Also, there is steeper increase in age groups 26 years to around 50 years, showing this age range is more frequently occurring group, consistent with previous observation

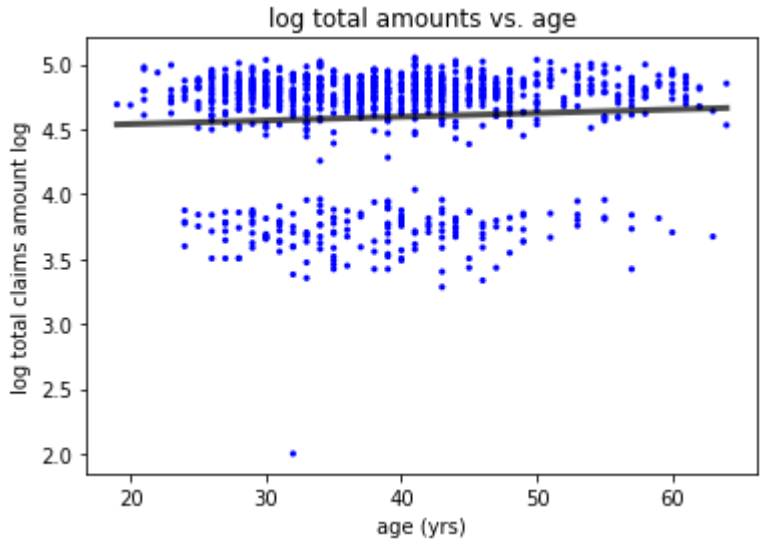
PLOT ANALYTICAL DISTRIBUTION OF VARIABLES. THE HIGHEST CORRELATED VARIABLES TO OUTCOME GLUCOSE AND BMI DISTRIBUTIONS ARE APPROXIMATELY GAUSSIANS OR NORMAL SO WE CAN APPLY MODEL

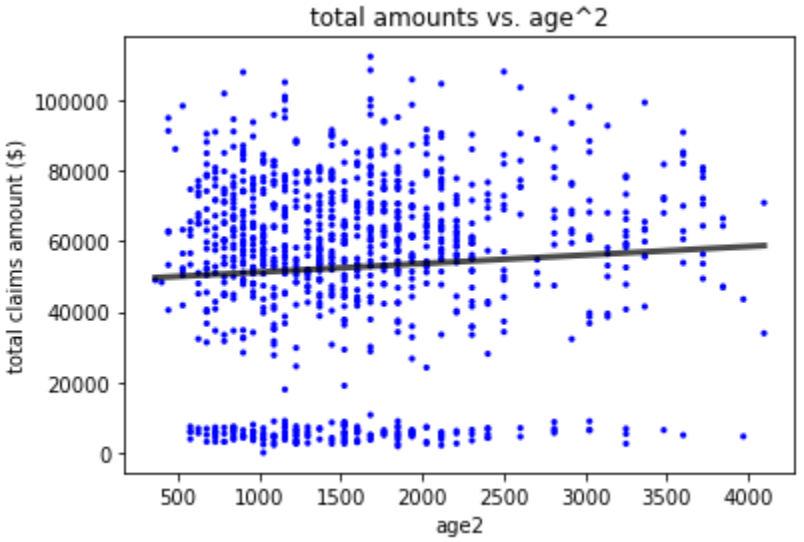


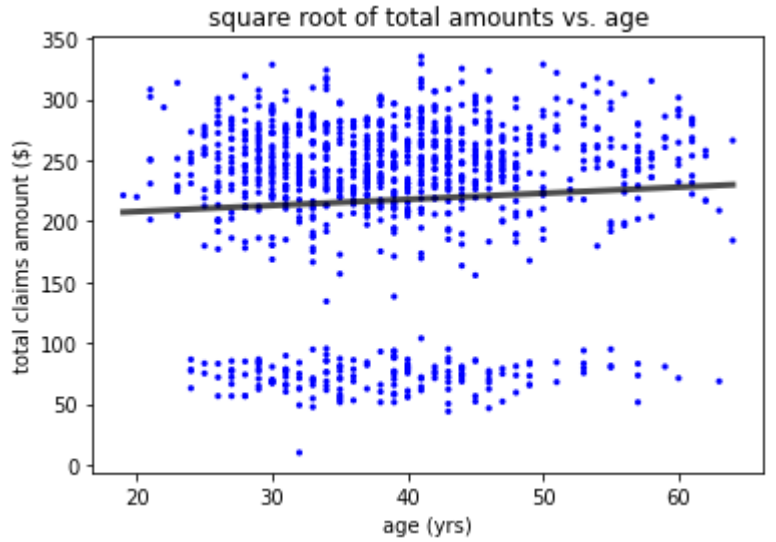
- The overall age group data for most accident-prone drivers falls within 2 standard deviations of the mean of 38.948 years i.e. 20.67 years to 57.29 years

SCATTERPLOT COMPARING VARIABLES - LINEAR AND NON-LINEAR RELATIONSHIP



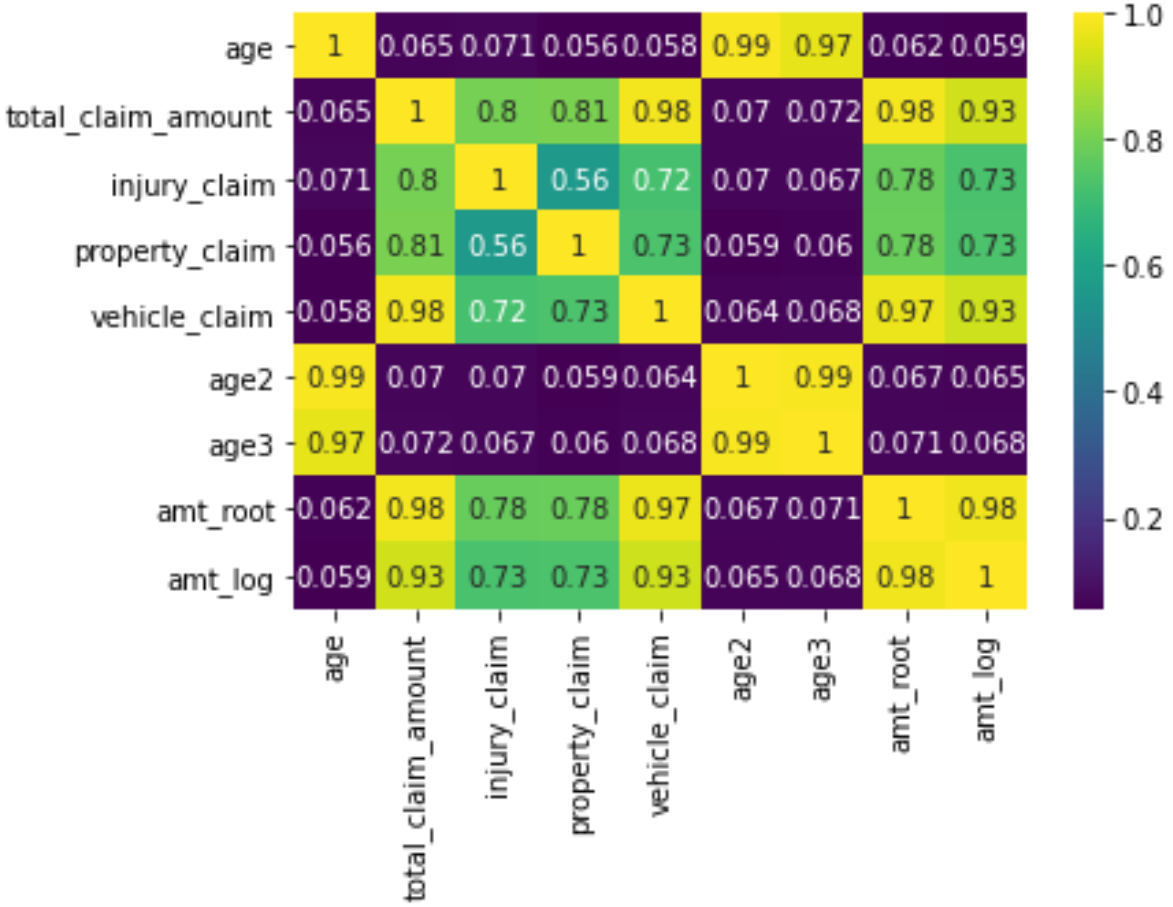






- Scatter plot of Total amount vs. age (Linear relation) – Top left corner
- Scatter plot of log10 of Total amount vs. age (Non-linear relation) – Top right corner
- Scatter plot of Total amount vs. age squared (Non-linear relation) – Lower left corner
- Scatter plot of square root of Total amount vs. age (Non-linear relation) – Lower right corner

CORRELATION ANALYSIS BETWEEN VARIABLES



- Total claim amount and age square seem to be correlated, though it is not a strong correlation (coefficient 0.07). Age and Age cube also have similar correlation (coefficient of 0.065 and 0.072) with total claim amount. Since all are close to each other, selecting any one of them might be okay for model creation.

HYPOTHESIS TESTING (USING AGE)

```
class CorrelationPermute(thinkstats2.HypothesisTest):

    def TestStatistic(self, data):
        xs, ys = data
        test_stat = abs(thinkstats2.Corr(xs, ys))
        return test_stat

    def RunModel(self):
        xs, ys = self.data
        xs = np.random.permutation(xs)
        return xs, ys
```

```
def CorTest1(n):

    m = 1000

    claims_df_subset = claims_df.dropna(subset=['age', 'total_claim_amount'])

    data = claims_df.age.values[:n], claims_df.total_claim_amount.values[:n]

    ht = CorrelationPermute(data)
    pvalue = ht.PValue(m)

    return round(pvalue, 3)
```

```
print(f"p-value for driver's age vs total claim amount correlation for 999 records: ", CorTest1(999))
```

```
p-value for driver's age vs total claim amount correlation for 999 records:  0.046
```

Looking at above pvalue, we can see the correlation between age and total claim amount is around 0.04 i.e. less than 0.05 and hence it looks to be Statistically Significant

- at p-value<0.05, the variable influences the outcome. By chance, it is unlikely that age variable does not affect the total claim amount and chance of accident. So age as a variable is statistically significant.

We can reject the Null Hypothesis and say that glucose level has a high chance of predicting an outcome of diabetes

HYPOTHESIS TESTING (USING AGE^2)

```
def CorTest2(n):

    m = 1000

    claims_df_subset = claims_df.dropna(subset=['age', 'total_claim_amount'])

    data = claims_df.age2.values[:n], claims_df.total_claim_amount.values[:n]

    ht = CorrelationPermute(data)
    pvalue = ht.PValue(m)

    return round(pvalue, 3)
```

```
print(f"p-value for driver's age squared vs total claim amount correlation for 999 records: : ", CorTest2(999))
```

```
p-value for driver's age squared vs total claim amount correlation for 999 records: :  0.022
```

Looking at above pvalue, we can see the correlation between age and total claim amount is around 0.02 / 0.03 i.e. less than 0.05 and hence it looks to be Statistically Significant

- at p-value<0.05, the variable influences the outcome. By chance, it is unlikely that age variable does not affect the total claim amount and chance of accident. So age as a variable is statistically significant.

We can reject the Null Hypothesis and say that glucose level has a high chance of predicting an outcome of diabetes

REGRESSION ANALYSIS

Dep. Variable:	total_claim_amount	R-squared:	0.067
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	1.023
Date:	Sat, 21 Nov 2020	Prob (F-statistic):	0.430
Time:	22:22:28	Log-Likelihood:	-11551.
No. Observations:	999	AIC:	2.323e+04
Df Residuals:	933	BIC:	2.356e+04
Df Model:	65		
Covariance Type:	nonrobust		

REGRESSION ANALYSIS (CONTINUED…)

Looking at the outcome from OLS method from statsmodel (statistics embedded within jupyter notebook. Due to large size, could not copy in everything), we can notice that

- Variables like age2 (age squared), auto make (e.g. Dodge, Ford) and auto model (e.g. E400, Escape) to an extent, insured occupation (e.g. transport moving, machine op inspector, handlers-cleaners) to some extent and insured education (e.g. MD, PhD etc) to some extent do have statistical significance in their respective relation with total claim amount field.

- On the other hand, variables like home state vs. accident state, insured sex, insured relationship did not seem to show much statistical significance in their respective relation with total claim amount field.

- Durbin Watson test statistics of 1.925 which is very close to 2 and shows that there is very little / no auto correlation between independent variables selected for the model.

- Skew is -0.564 which indicates that data is negatively skewed i.e. longer tail in the data on the lower end of the distribution. This also points back to the observation from the Histogram for Total Claim Amount that the amounts less than $20,000 were appearing as a skew

- Kurtosis value of 2.617, which is greater than 1 indicates that the distribution is peaked i.e. pointy in shape