1. **Insights from data.**

Histogram

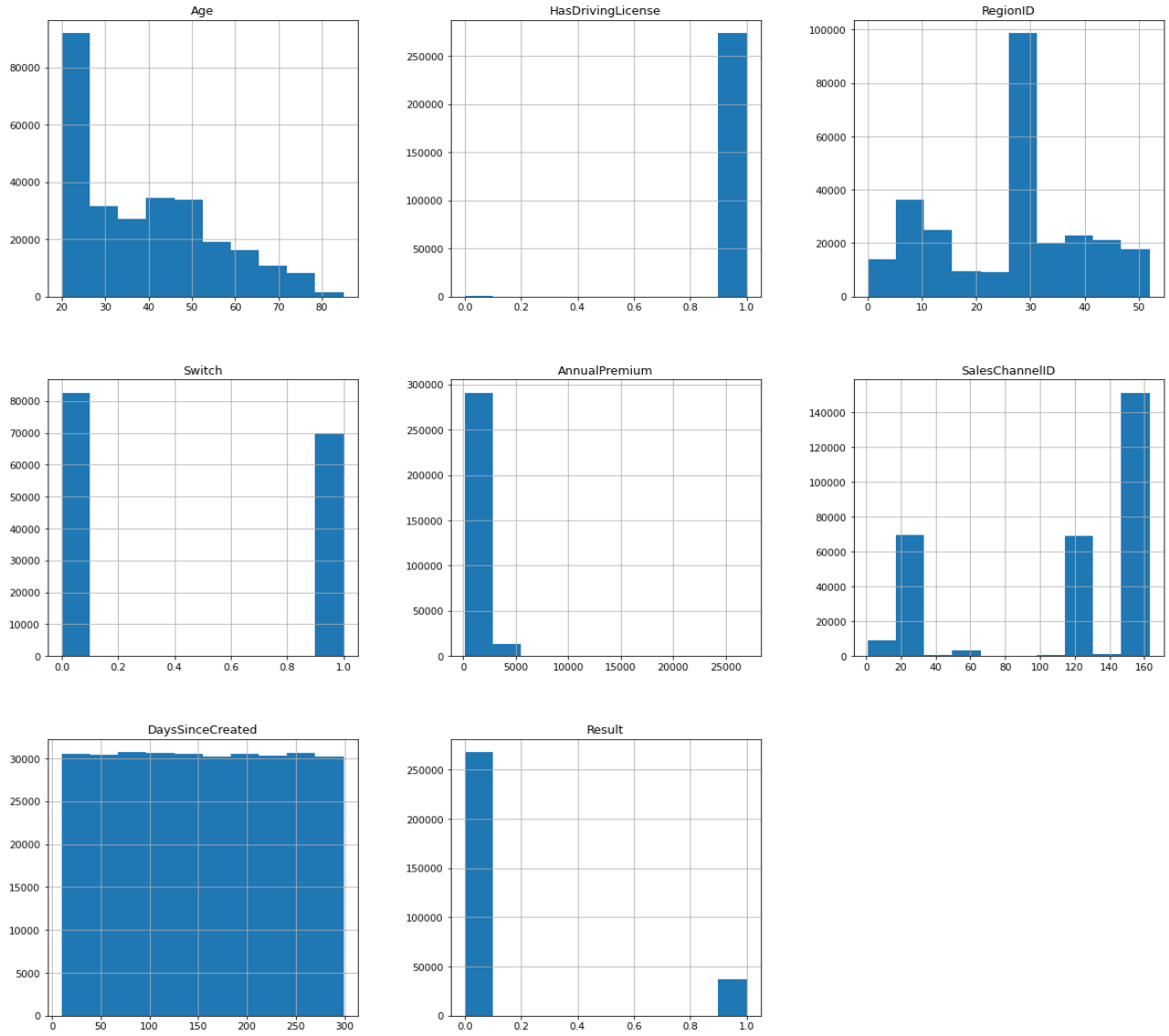


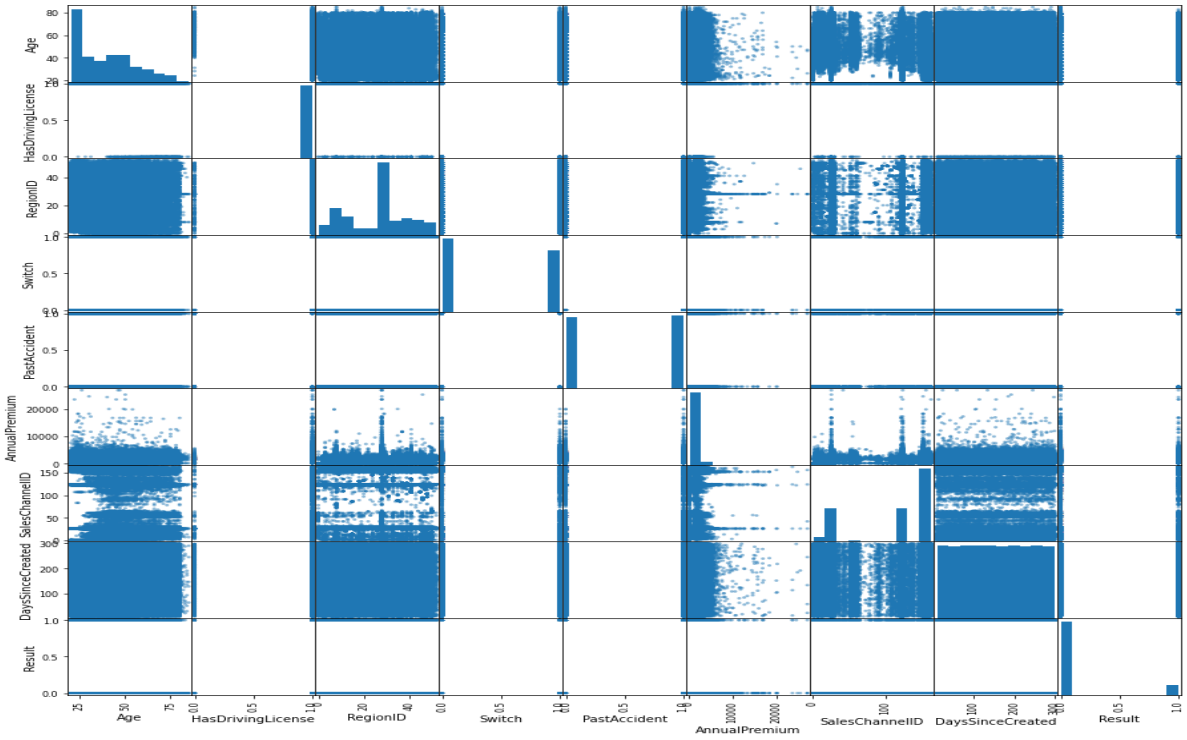
Figure 1: Histogram showing trends in data.

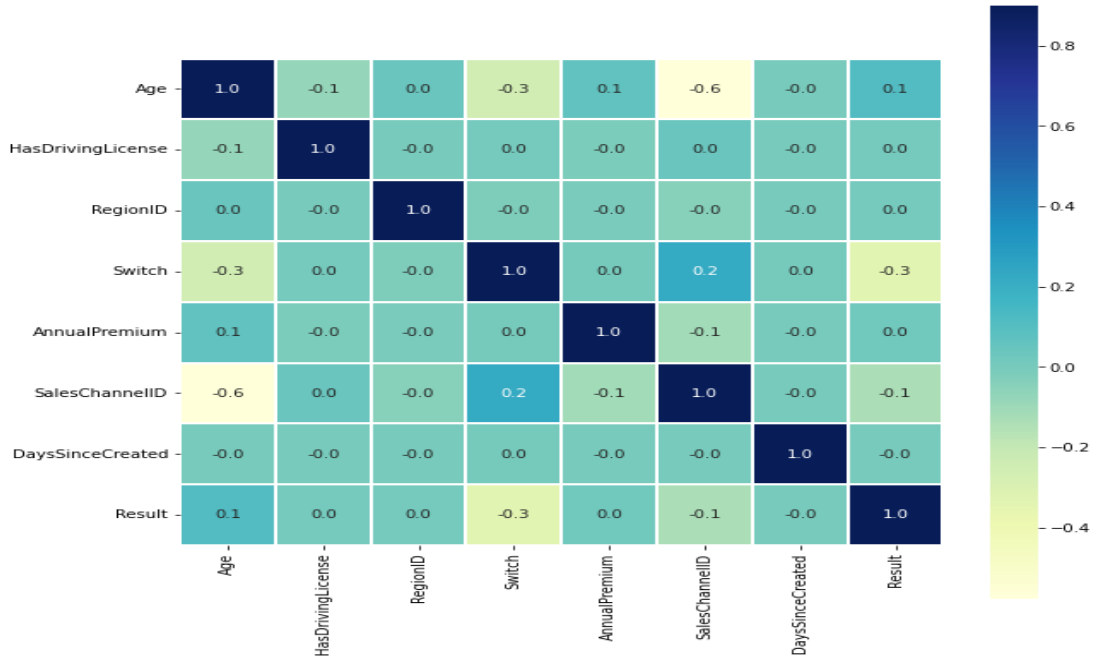
Key Takeaways:

- Majority of people are between Age 30-50 years. While people between 20-25 years are highest.
- Almost all of them have a license
- There is a specific regionID between 25-30 from which people belongs most.
- People switching insurance provider and not switiching are relatively balanced

- 75% of people pay Annual premium=1970 GBP which is greater than mean Annual premium of 1528 GBP.
- Majority of sales lead are found in between 145-160 Sales Channel ID.
- People buying insurance is uniformly distributed over the past 1 year.
- Number of people that are buying car insurance is drastically smaller than people not buying it.

Key takeaways from Scatterplot and Correlation matrix





- There is -0.6 correlation between Age and SalesChannelID. Indicating older people prefer lower SalesChannelID or Young people prefer higher SalesChannelID.

- Also, with -0.3 correlation between Age and Switch, it indicates older people doesn't prefer to switch car insurance provider or Young people prefer to switch car insurance provider more frequently.

- With Switch and Result having -0.3 correlation it indicates those people switching car insurance has a lower probability of buying a car insurance.

- With SalesChannelID and Switch having 0.2 correlation, it indicates people preferring higher SalesChannelID i.e. Young people (20-25 years) change their car insurance provider more often.

2. Machine learning methods selection and implementation

Our goal here is to correctly classify the Result as 1 or 0 based on its input classes of dataset, also we will identify the algorithm which is best suited for this problem.

• **Since we want to do classification (Result: 1/0**) we apply various classification models and do hyperparameter tuning. We use, Logistic Regression, XGB Classifier, Decision Tree, SVM Classifier, Random Forest Classifier, ADA Boost Classifier, and Gradient Boost Classifier. Then we select the best 3 performers.

Choosing 3 Models based on accuracy score and model generalisation capabilities.

1. LogisticRegression(): Best Accuracy : 87.87%

2. SVC(): Best Accuracy : 87.87%

AdaBoostClassifier,GradientBoostingClassifier and XGBClassifier all have accuracy= 87.87%. Also, we have outliers and Gradient Boosting algorithm is more robust to outliers than AdaBoost, we choose Gradient Boosting over AdaBoost. Further, XGBoost uses advanced regularization (L1 & L2), which improves model generalization capabilities. XGBoost delivers high performance as compared to Gradient Boosting. Also, the imbalance in

our target variable is well adjusted by XGBoost and hence we choose XGBoost between all of them.

3. XGBoost

With GridSearchCV we loop through predefined hyperparameters and fit our estimator (model) on our training set. So, in the end, I can select the best parameters from the listed hyperparameters.

3. Model Performance Evaluation

1. Logistic Regression

```
precision    recall  f1-score   support

           0       0.88      1.00      0.94     20048
           1       0.00      0.00      0.00      2767

    accuracy                           0.88     22815
   macro avg       0.44      0.50      0.47     22815
weighted avg       0.77      0.88      0.82     22815

ROC AUC score: 0.8182757839443376
Accuracy Score:  0.8787201402586018
```

2. XGBoost

```
precision    recall  f1-score   support

           0       0.88      1.00      0.94     20048
           1       1.00      0.00      0.00      2767

    accuracy                           0.88     22815
   macro avg       0.94      0.50      0.47     22815
weighted avg       0.89      0.88      0.82     22815

ROC AUC score: 0.8500538984716405
Accuracy Score:  0.8787639710716634
```

3. SVC

```
precision    recall  f1-score   support

           0       0.88      1.00      0.94     20048
           1       0.00      0.00      0.00      2767

    accuracy                           0.88     22815
   macro avg       0.44      0.50      0.47     22815
weighted avg       0.77      0.88      0.82     22815

ROC AUC score: 0.7793794441587389
Accuracy Score:  0.8786763094455402
```

4. Summary:

Model	Accuracy
1 XGBoost	**87.876397**
2 Logistic Regression	**87.872014**
3 SVC	**87.867631**

- After conducting significant data analysis, I experimented with various classification models to see how well they performed on the dataset. With accuracy, roc, precision, and recall score, I obtained quite decent results. I think choosing other methods for filling missing values in Switch and PastAccident columns can make Confusion matrix consistence.

- Using Grid Search, I fine-tuned the hyperparameters and did model evaluation using the classification report, which included Confusion matrix, ROC AUC, F1-score and Precision-Recall curves for various models.

- With that, I came to conclusion that Logistic Regression, SVC, and The Boosting Algorithm: XGBoost are models which are best fit for our dataset.

- After Optimization the XGBoost algorithm has the Highest Accuracy of exactly 87.87% & AUC of 0.85.