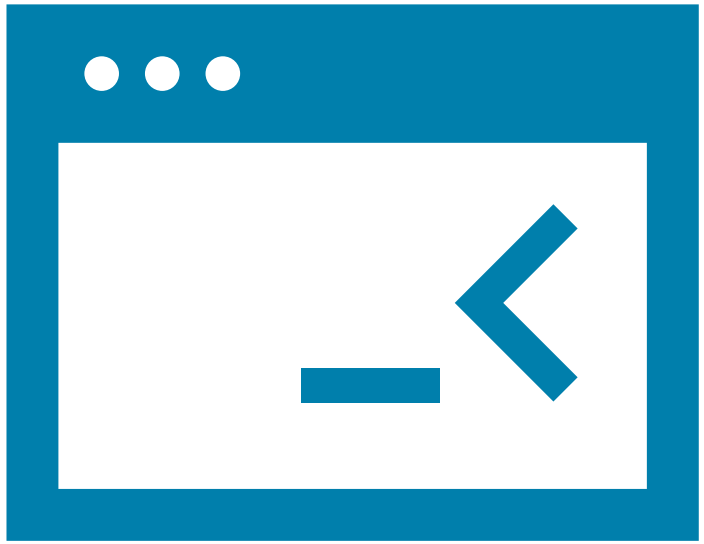
ANALYSIS OF THE EFFECT OF COVID ON THE PAYMENT OF HEALTHCARE EXPENSES AND PAYMENT PREDICTION USING MACHINE LEARNING ALGORITHMS

Capstone Presentation

RESEARCH QUESTION

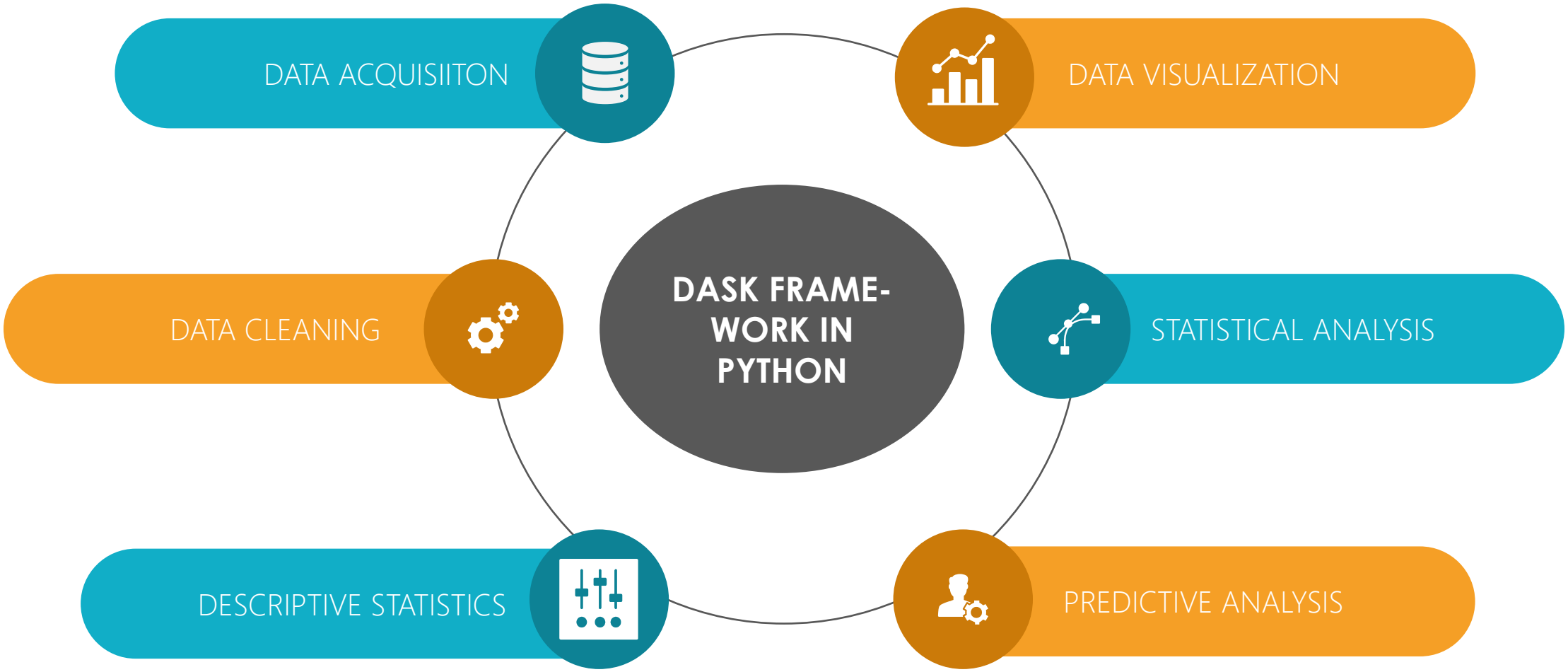
- This study analyses the types of insurances accepting or denying the insurance claim for the payment of healthcare expenses, the time taken for the payment, state wise analysis of the payments Pre-Covid and Post-Covid era. Which in turn helps to discover the effect of Covid on the payment of healthcare expenditures, and finally predict whether the unknown and future records would be paid or not. This study uses two datasets, firstly they are cleaned then goes on to calculation of descriptive statistics, data visualization, performing statistical tests such as t-test, ANOVA, Chi-SQ test, correlation test, including the alternate tests for each one of the tests. Finally, it uses a set of supervised machine learning models such as Logistic regression, linear regression, Gradient descent boost method, Random forests, Support vector machines, K-nearest neighbors. This study also evaluates the predictive power of supervised machine learning algorithms based on accuracy.



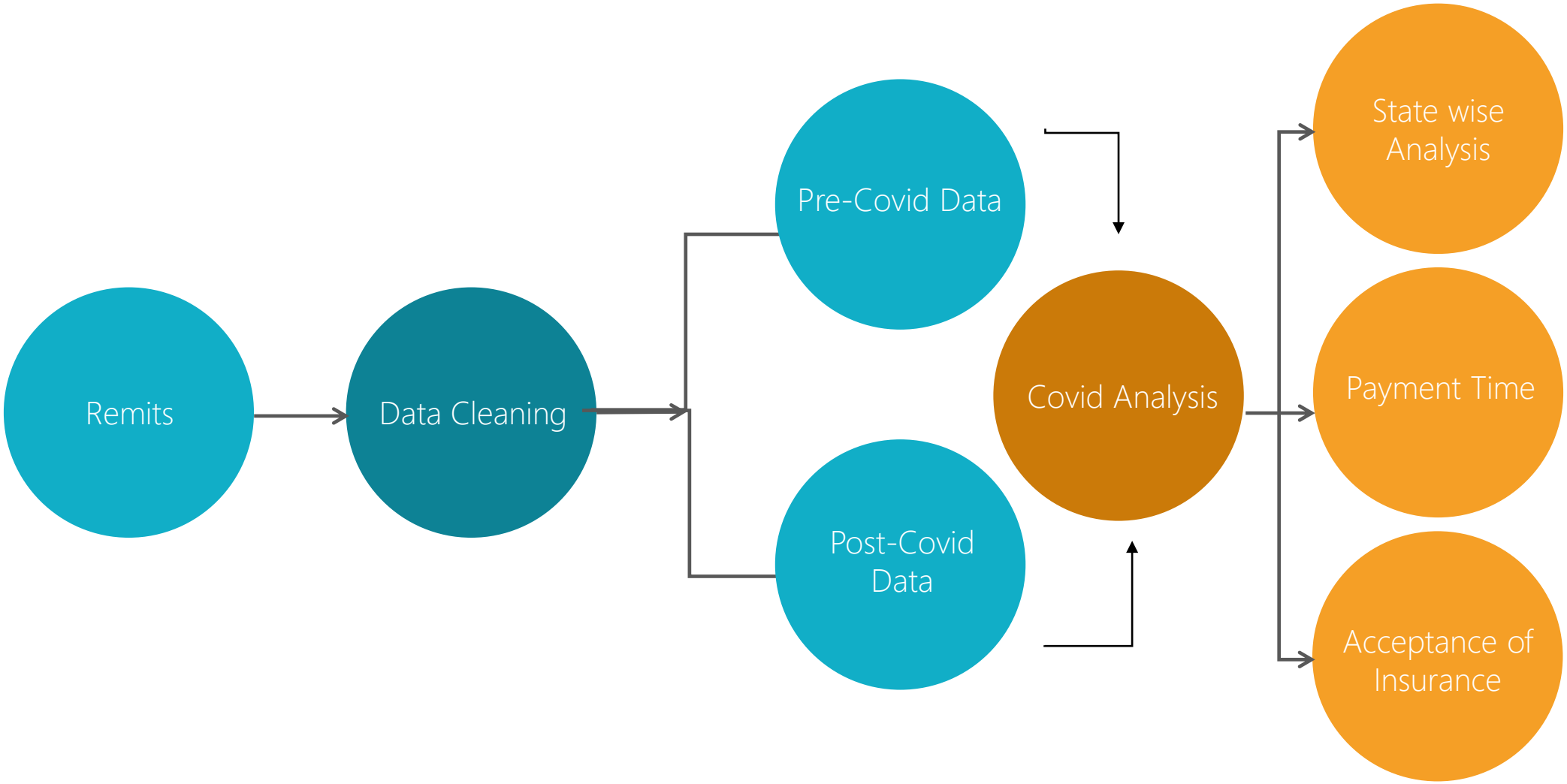
SOFTWARES USED

DASK FRAMEWORK IN PYTHON

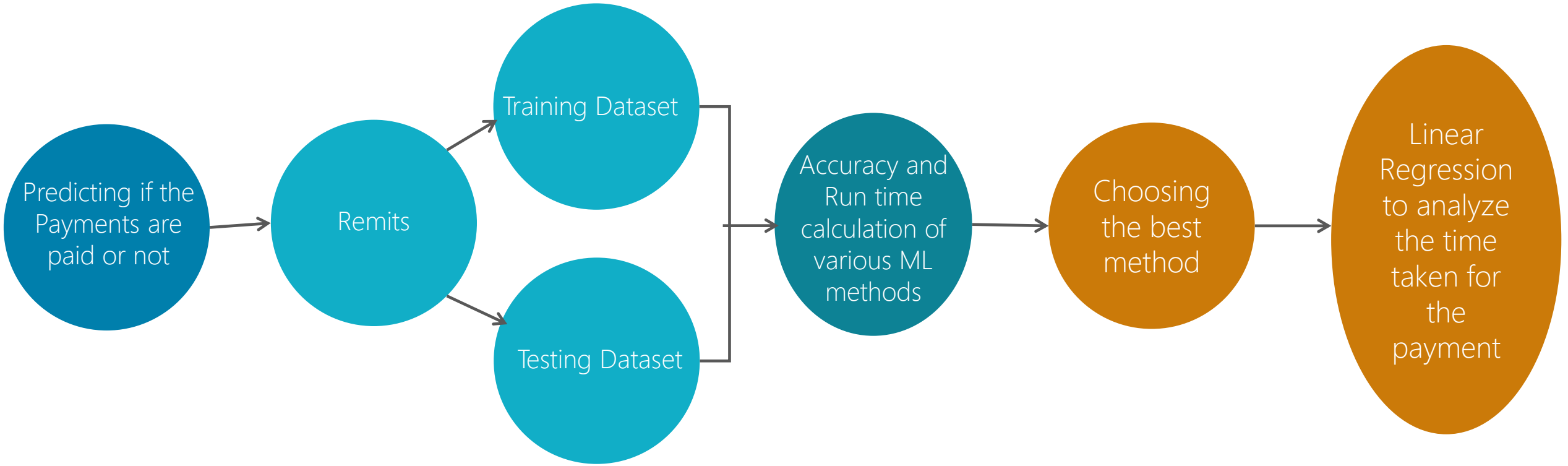
Project Road Map



Statistical Analysis



Predictive Analysis



Data Acquisition

- The data is taken from the data base in my company using a tool called VisiQuate in which I have filtered in the data based on the statement years 2018-2022

ATTRIBUTES USED

CLAIMS DATASET

- Statement End FY
- Denial Type
- Statement End Date
- Statement Start Date
- RMA_Payer: Prim Lvl III
- Claim Status Description
- Claim Status Category
- Claim Received Date
- Claim Filling Code
- Billing Provider State
- Billing Provider City
- Initial Denial Remit Count
- Total Denied Remit Count
- Final Denials Remit Count

REMITS DATASET

- Statement End FY
- Statement Start Date
- Statement End Date
- RMA_Payer: Prim Lvl III
- Admit Type
- Claim Bill Date
- Billing Provider State
- Billing Provider City
- Claim Count

EXPLORATORY DATA ANALYSIS

Data Cleaning

The nulls are removed from the dataset, the data type of the attributes are changed to the appropriate data types.

In Remits Data set the attributes Statement End Date, Statement Start Date, Claim Received Date are changed to datetime. The attributes Claim Status Category, RMA_Payer: Prim Lvl III, Denial Type, Claim Status Description, Claim Status Category, Billing Provider State, Billing Provider City are changed to categorical data type.

In claims data set the attributes Statement End Date, Statement Start Date, Claim Bill Date are changed to datetime. The attributes Billing Provider City, RMA_Payer: Prim Lvl III, Billing Provider State are changed to categorical data type.

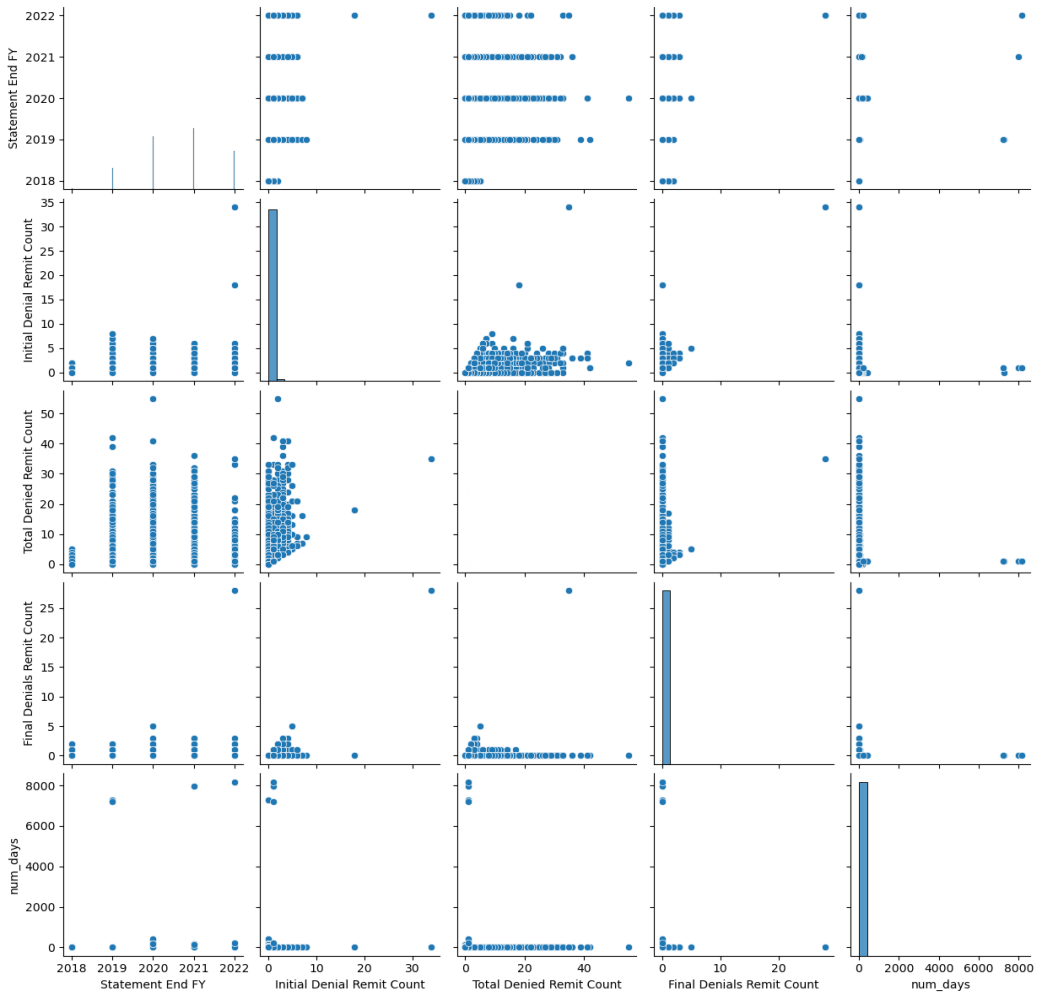
Descriptive Statistics

- The Descriptive statistics such as count mean, standard deviation, minimum value, maximum value, inter quartiles are calculated for the numerical attributes in both the remits (Statement End FY, Initial Denial Remit Count, Total Denied Remit Count, Final Denials Remit Count) and claims dataset (Statement End FY, Claim Count).
- There are 516964 total records in the remits dataset, where initially 34 remits are denied at maximum and 28 the final denial remit count. However, most of the claim counts are received after the third quartile (75%). The maximum total remits count is 55. The maximum number of days needed to complete the payment of claims is 8172. The results are tabulate in Table 1. The maximum number of claims received from insurance companies is 52 claims in the claims data set.

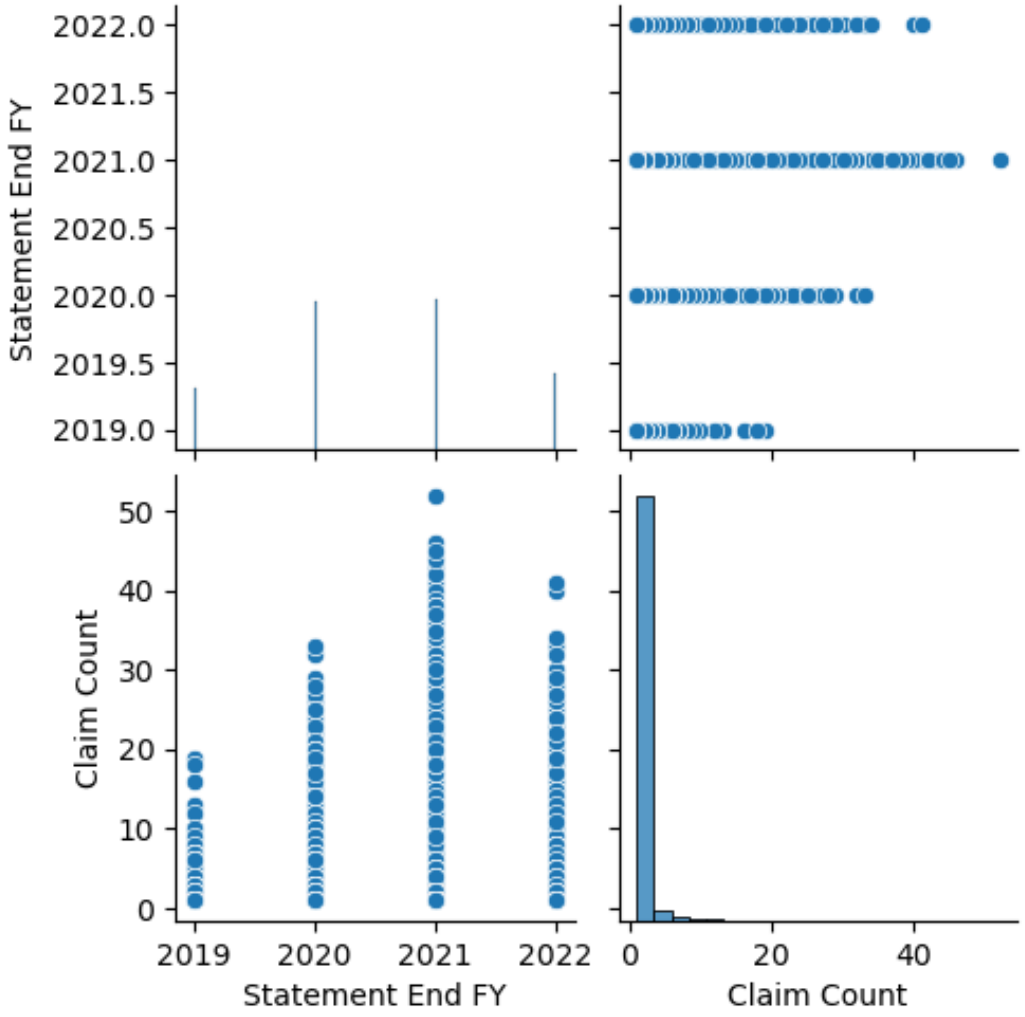
DATA VISUALIZATION

- The visualizations namely scatter plots, heatmaps, stacked bar charts, and bar charts on both the datasets i.e., Claims and remits datasets. The data visualization charts has revealed interactions between RMA_Payer: Prim Lvl III and Claim Status Category, as well as RMA_Payer: Prim Lvl III and Billing Provider State. Scatter plots revealed the extent linearity between the numerical variables, and heatmaps have revealed the extent of correlation.

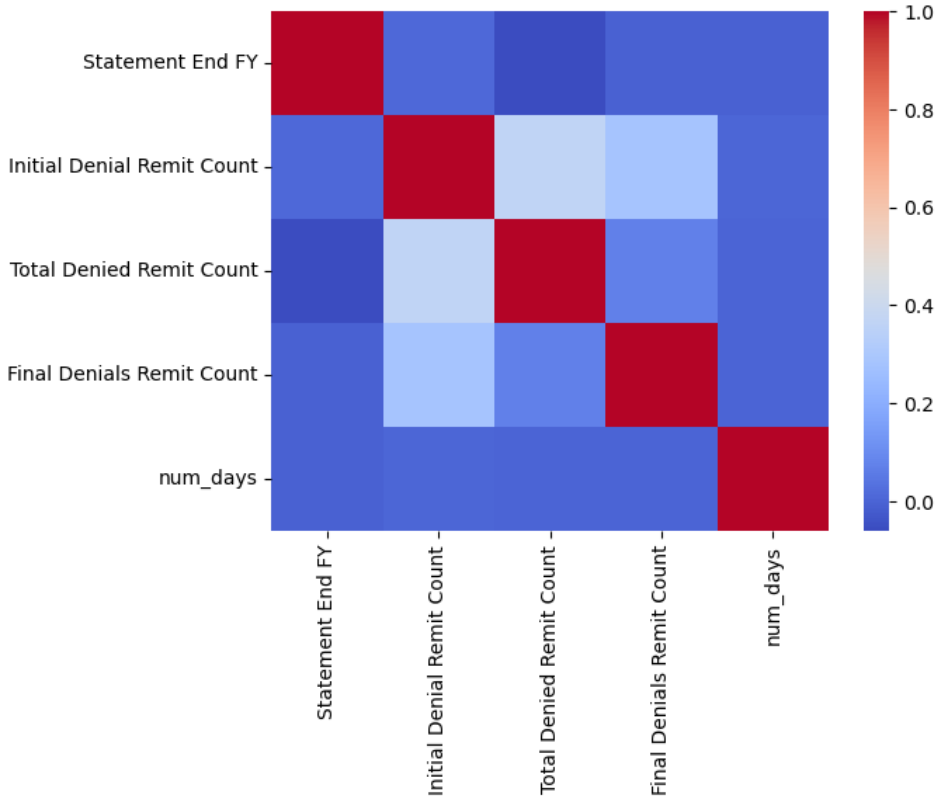
Scatter Plots



Remits Data Set



Claims Data Set

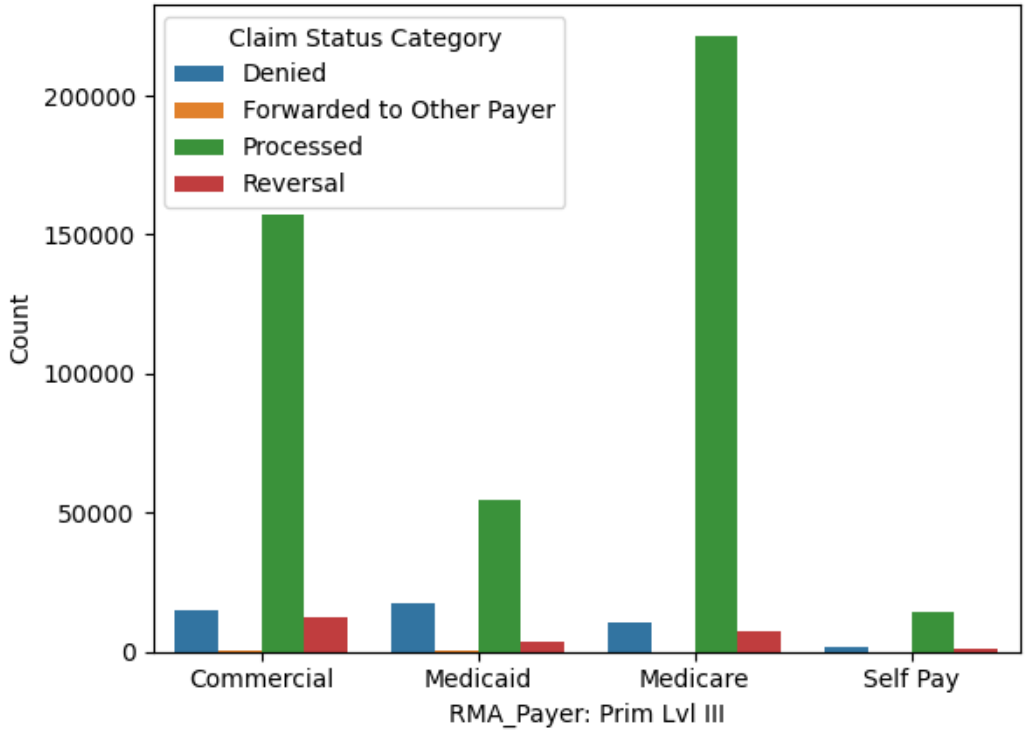


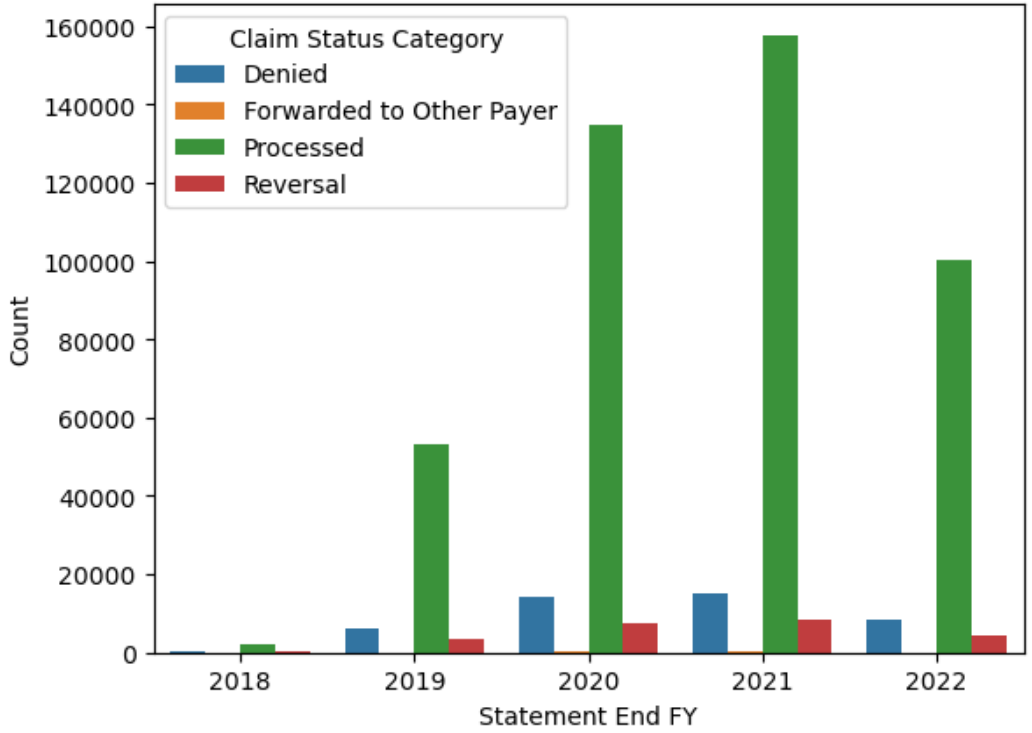
Remits Data

Claims Data

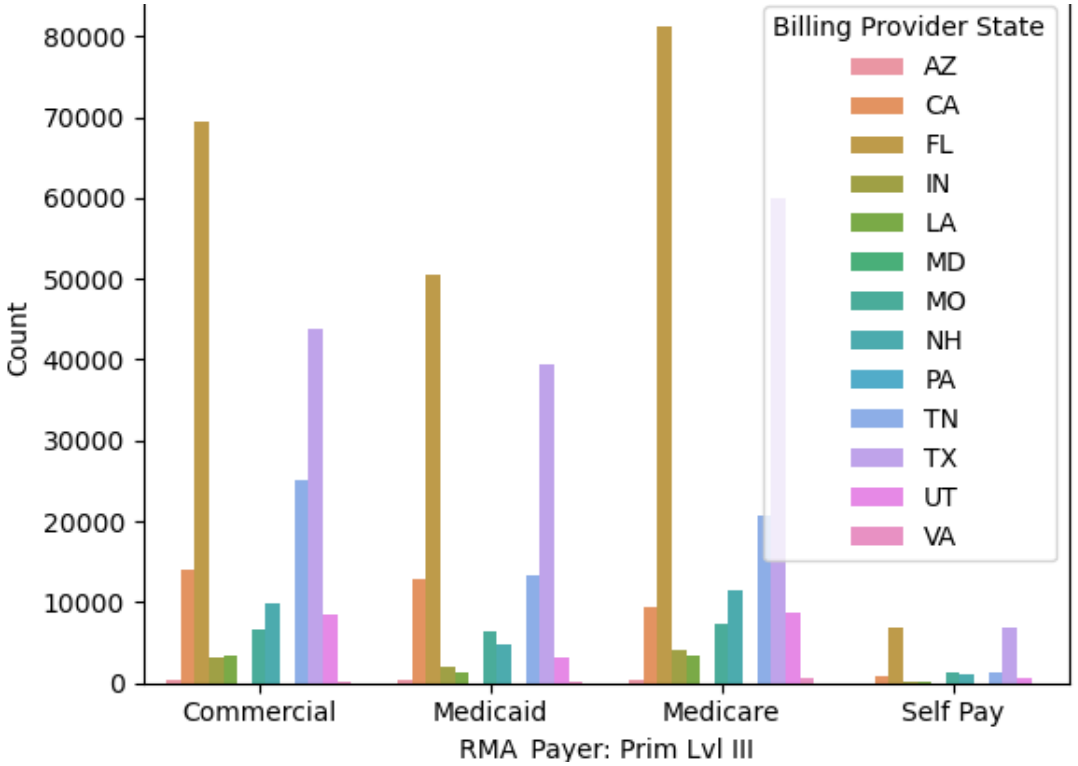
Correlation Plots

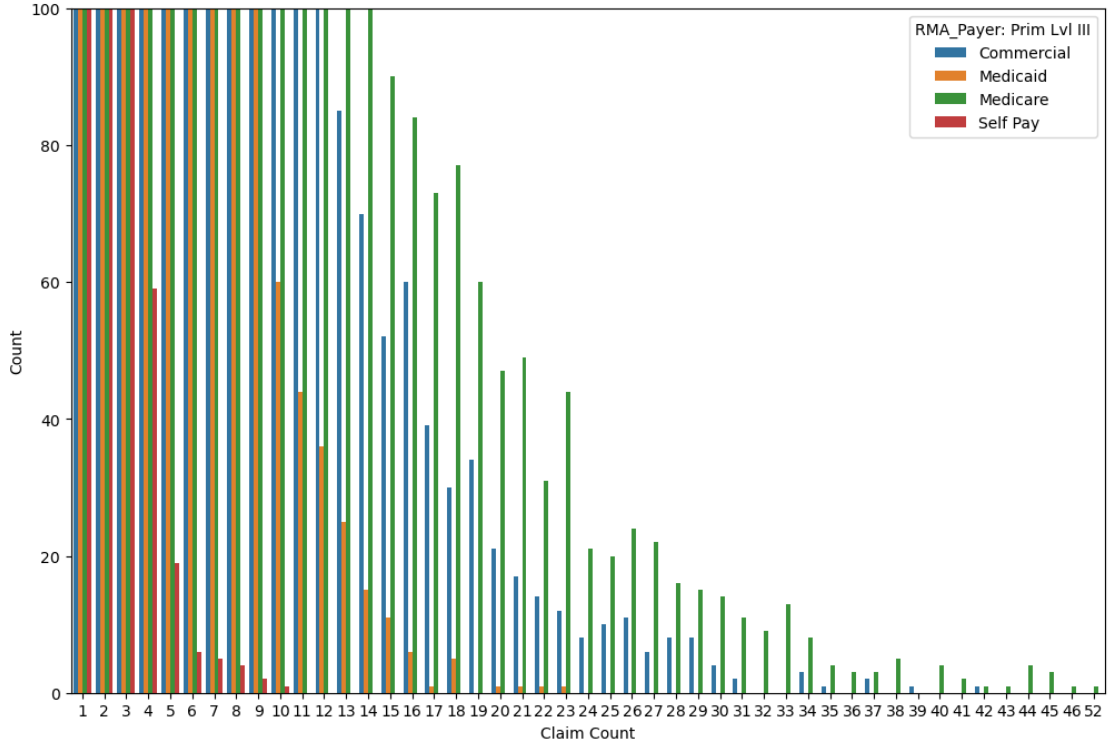
Stacked Bar Charts for Remits Data





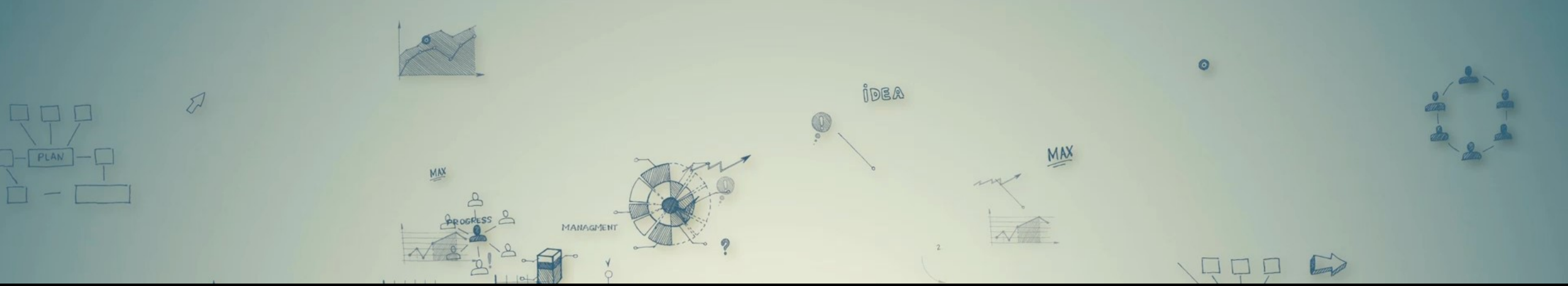
Stacked Charts for Claims Data





While visualizing the remits dataset, it is seen that most of the records in the data set are Medicaid payers and very few are from self-pay. From the scatter plots it is found that there is no linearity between the number of days needed for the payment and the denial remit count, and there is a correlation between the denial remit count. From the correlation plots here is a slight correlation between year and initial denial remit count. From the stacked bar plot, it is observed that most of the Medicare payments are processed i.e., the claims are retrieved. Among all the payers, Medicaid insurances are the most denied ones. Among all the payers the commercial payers had more claims which are reversed. None or very few claims are forwarded to other payers among all types of payers. It is also found that most of the processed claims are in the year 2021 and the least are in 2018. Most of the claims are denied in the year 2021, and very few claims are denied in 2018.

In the claims dataset, medicare had more records and less self-pay. From scatter plots it is observed that there is no linearity between the claim count and year. From the stacked bar charts, among all other states, Florida has the most records in all the payer types and among those Medicare has the most records. Other states having the most records in all payers is Texas and Tennessee. It is also observed that most of the payers had a count below 10 claims. Moreover, from heatmaps it is clear that there is no correlation between the claim count and year.


PLAN
MAX
PROGRESS
MANAGMENT
IDEA
MAX


OPTION
START
IDEA
up
down
IDEA!!!

To analyze the trends in pre-covid and post-covid era, the remits dataset is divided into two parts based on the date. The records before 01-01-2020 are included in the pre covid dataset and the records after 01-01-2020.

Analysis of types of insurances accepting or denying the claims

Firstly, chi-squared test is performed to know whether there is a relationship between the type of insurance and the claim status category The outcome variable is RMA_Payer: Prim Lvl III) and the predictor variable is claim status category, both are categorical variables. Also, the assumptions of chi-squared test namely, all observations should be independent, all the variables should be nominal or ordinal, and the expected values should be 5 or higher in at least 80% of groups are checked. Then the percentages of denied and accepted claims for each type of insurance.

Chi-Squared Test

- From the chi-square test it is revealed that there is a relationship between Claim Status Category and the RMA_Payer: Prim Lvl III as the p-value is less than 0.05. When the standardized residuals is performed to check which cell is contributing towards the chi-squared statistic, it is observed that all the standardized residual values are between -2 and 2 indicating that all cells in the contingency table is not contributing significantly to the overall chi-square test statistic. The odds ratios equal to 1 reveal that there is slightly a positive association between all the groups of claims status category and RMA_Payer: Prim Lvl III.
- All the assumptions of the chi-squared test are met. Specifically, since all the observations are collected independently, the assumption all observations are independent is met. As the observations are measured as nominal, the assumption of all the variable should be nominal or ordinal is met. The assumption of the expected values should be 5 or higher in at least 80% of groups is met. These results are the same throughout the years i.e., there is no fluctuations in the pre-covid and post-covid Era.

Percentages of denied and accepted claims for each type of insurance.

the denial of commercial claims increased by 2% but the self-pay denial decreased by 10%. Interestingly the payment trend was positive during the post-covid period.

Pre-COVID era

RMA_Payer: Prim Lvl III	Claim Status Category	Proportion Change
Commercial	Denied	6.274348
	Forwarded to Other Payer	0.105413
	Processed	85.384298
	Reversal	8.235941
Medicaid	Denied	33.509458
	Forwarded to Other Payer	0.021135
	Processed	59.019338
	Reversal	7.450069
Medicare	Denied	4.847611
	Forwarded to Other Payer	0.003066
	Processed	91.773472
	Reversal	3.375851
Self-Pay	Denied	20.595383
	Forwarded to Other Payer	0.000000
	Processed	69.623329
	Reversal	9.781288

Post-COVID era

RMA_Payer: Prim Lvl III	Claim Status Category	Proportion Change
Commercial	Denied	8.297133
	Forwarded to Other Payer	0.135828
	Processed	84.963473
	Reversal	6.603566
Medicaid	Denied	21.070087
	Forwarded to Other Payer	0.189302
	Processed	74.094378
	Reversal	4.646232
Medicare	Denied	4.249054
	Forwarded to Other Payer	0.005811
	Processed	92.889832
	Reversal	2.855303
Self-Pay	Denied	10.265372
	Forwarded to Other Payer	0.019417
	Processed	85.682848
	Reversal	4.032362

Analysis of time taken for the payment based on the payer type

The Multiple Linear Regression is performed to analyze whether there is a relationship between the type of insurance and the number of days. The outcome variable is RMA_Payer: Prim Lvl III and the predictor variable is number of days. Also, the assumptions of multinomial regression namely independence of observations, linearity, no perfect multicollinearity are checked.

Multinomial Regression

- From the Multinomial regression it is revealed that there is a change in the log of odds of the outcome and the p-value in the post-covid era compared to its counterpart. During the pre-covid era, the coefficients of Medicaid, Medicare, Self-pay are negative, indicating a decrease in the odds of the outcome. The p-value is more than 0.05, so coefficient is not statistically significant, and the AIC value reveals that the model is not a good model. Also, statistics reveal that the Medicaid claims took less time to get paid, the commercial and Medicare plans took more time. During Post-covid era, the coefficients of Medicaid and Medicare are positive indicating an increase in the odds of the outcome, and the Self-Pay coefficient is negative, indicating a decrease in the odds of the outcome. The p-value is less than 0.05 for the Medicare claims, so the coefficient is statistically significant and the p-value for Self-Pay is less than 0.05, so the coefficient is not significant.
- The AIC value reveals that the model is not a good model, and the statistics reveal that the Medicaid Claims took less time to get paid, however Medicaid claims took more time to get paid. All the assumptions of the multinomial regression are met except the linearity assumption for both pre-covid and post-covid dataset. Overall, the Number of days taking for a claim to get paid decreased in the post-covid era, The Medicaid and Medicare insurance types became statistically significant post-covid in regards of the time taken for the payment.

Analyzing the claim status and number of days needed for claim reimbursement based on state

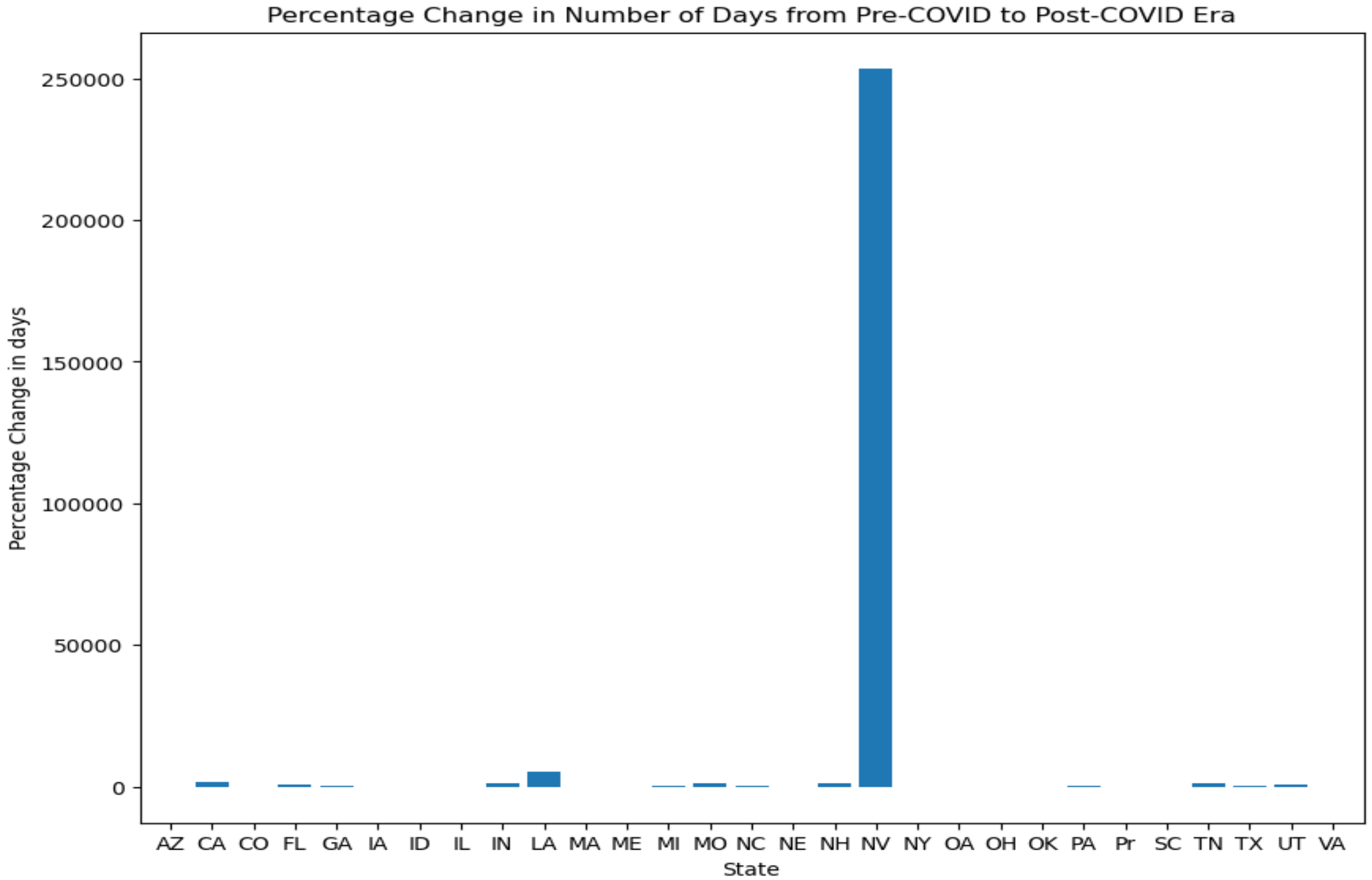
To perform state wise analysis, the chi-squared test is performed to analyze the relationship between Claim Status Category and Billing Provider State. Also, the assumptions of chi-squared test namely, all observations should be independent, all the variables should be nominal or ordinal, and the expected values should be 5 or higher in at least 80% of groups are checked. Then the effect of state on the claim status and percentage change in number of days from pre-COVID to post-COVID are visualized.

Chi- Squared Test

- From the chi-squared test, it is observed that during the pre-covid and post-covid era, there is a relationship between as the p-value is less than 0.05. The odds ratios for both the eras is 1 revealing that there is a slightly positive association between all the groups of claims status category and Billing Provider State. However, when the standardized residuals is performed to check which cell is contributing towards the chi-squared statistic, it is observed that for denied, processed, and reversal claims during pre-covid era are between -2 and 2 indicating that all cells in the contingency table is not contributing significantly to the overall chi-square test statistic. The standardized residual values for all states except Indiana, Louisiana, Michigan, Tennessee, Utah are less than -2, so they are contributing significantly to overall chi-square test statistic. The standardized residual values during post-covid era for denied, processed, and reversal claims are between -2 and 2 indicating that all cells in the contingency table is not contributing significantly to the overall chi-square test statistic. The standardized residual values for all states except California, Florida, Louisiana, Missouri, New Hampshire, Nevada, South Carolina are less than -2, so are contributing significantly to overall chi-square test statistic.
- For the pre-covid era, all the assumptions of the chi-squared test are met except one. Specifically, since all the observations are collected independently, the assumption all observations are independent is met. As the observations are measured as nominal, the assumption of all the variable should be nominal or ordinal is met. The assumption of the expected values should be 5 or higher in at least 80% of groups is not met.

Statistics and Visualization

It is observed that during pre-covid era Pennsylvania took more time to complete the payment throughout all status. New Hampshire, Nevada, and New York took the least time to complete the payment. However, during the post-covid era Indiana took more time to make the payment, Colorado, Iowa, Idaho, Illinois, Maine, Nebraska, Ohio, Oklahoma, South Carolina, Virginia, Puerto Rico took very less time. From the visualizations it is observed that the percentage change in number of days taken to complete the payment is larger in Nevada, and there is no change in South Carolina, Puerto Rico, Oklahoma, Ohio, New York, Nebraska, Manie, Massachusetts, Illinois, Idaho, Iowa, Colorado, and Arizona (Fig 5). Overall, there is a relationship between the Billing Provider State and the Claim Status Category, and the payment time is improved post covid.


Percentage Change in Number of Days from Pre-COVID to Post-COVID Era
Percentage Change in days
250000
200000
150000
100000
50000
0
AZ CA CO FL GA IA ID IL IN LA MA ME MI MO NC NE NH NV NY OA OH OK PA Pr SC TN TX UT VA
State


Predictive Analysis

Predicting if the claims are paid or not

Various supervised machine learning classification models namely, Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors, Support Vector Machines, Random Forest, Gradient Boost, XG Boost are used to predict if the claims are paid or not as the outcome variable is categorical. Accuracy and Time is also computed for each of the models to identify the best model for predicting the payment of claims, the data set used is the remits data set. The outcome variable in the predictive analysis is Claim Status Category and the predictor variables are Statement End FY, RMA_Payer: Prim Lvl III, Claim Status Description, Denial Type, Claim Filling Code, Billing Provider State, Billing Provider City, Initial Denial Remit Count, Total Denied Remit Count, Final Denials Remit Count, num_days. The outcome variable is converted to binary variable encoding the denied category as 1 and rest as 0. The predictor variables are also label encoded before fitting the models.

The best model is Gaussian Naive Bayes as it took less time and provided 100 percent accuracy. The other better models are logistic regression, XG boost, Random Forest, Gradient Boost respectively

Machine Learning Algorithm	Time in seconds	Accuracy in percentage	Rank
Logistic Regression	3.42	100	2
Gaussian Naive Bayes	0.46	100	1
K-Nearest Neighbors	87.31	99.8	6
Support Vector Machines	960.21	99.9	7
Random Forest	17.33	100	4
Gradient Boost	21.85	100	5
Xg Boost	5.74	100	3

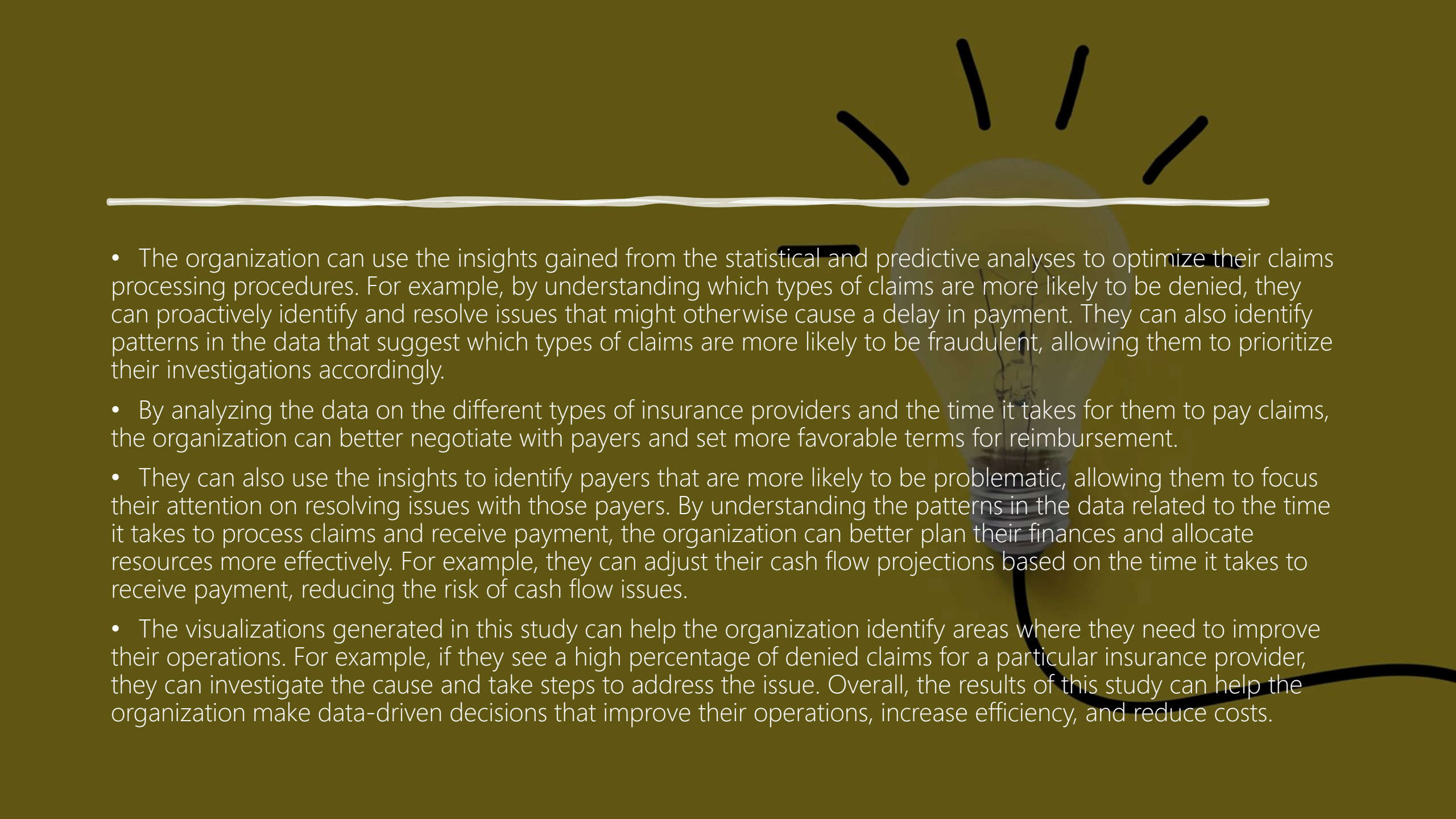
Predicting time taken to make the entire payment

A linear regression model is fit to predict the time taken for the payment as the outcome variable (num_days) here is numerical. The model is first fit on the remits data set and the score is calculated then the same model is applied on the claims data set considering it as the new records. The predictor variables are RMA_Payer: Prim Lvl III, Billing Provider State, Billing Provider City and the outcome variable is num_days.

SUMMARY

The study used the Dask framework in Python and claims and denial data from the Envision Healthcare database to analyze trends in pre-COVID and post-COVID era. The datasets were cleaned by removing nulls and changing data types to appropriate types. Descriptive statistics were calculated for numerical attributes in both datasets, and visualizations like scatter plots, heatmaps, stacked bar charts, and bar charts were used to reveal interactions between variables. Statistical tests like Chi-Squared test and Multinomial regression were used to analyze the datasets, and machine learning classification models like Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors, Support Vector Machines, Random Forest, Gradient Boost, XG Boost were used to predict if claims were paid or not. A linear regression model was also used to predict the time taken for the payment. The study found a relationship between the type of insurance and claim status category, and that there was a relationship between the type of insurance and the number of days taken for payment. The study also found a relationship between claim status category and billing provider state, with a percentage change in the number of days from pre-COVID to post-COVID. In the predictive analysis it is found that the best model for predicting payment of claims was Random Forest, and the best model for predicting the time taken for payment was Linear Regression.

for organizatior



- The organization can use the insights gained from the statistical and predictive analyses to optimize their claims processing procedures. For example, by understanding which types of claims are more likely to be denied, they can proactively identify and resolve issues that might otherwise cause a delay in payment. They can also identify patterns in the data that suggest which types of claims are more likely to be fraudulent, allowing them to prioritize their investigations accordingly.
- By analyzing the data on the different types of insurance providers and the time it takes for them to pay claims, the organization can better negotiate with payers and set more favorable terms for reimbursement.
- They can also use the insights to identify payers that are more likely to be problematic, allowing them to focus their attention on resolving issues with those payers. By understanding the patterns in the data related to the time it takes to process claims and receive payment, the organization can better plan their finances and allocate resources more effectively. For example, they can adjust their cash flow projections based on the time it takes to receive payment, reducing the risk of cash flow issues.
- The visualizations generated in this study can help the organization identify areas where they need to improve their operations. For example, if they see a high percentage of denied claims for a particular insurance provider, they can investigate the cause and take steps to address the issue. Overall, the results of this study can help the organization make data-driven decisions that improve their operations, increase efficiency, and reduce costs.

conclusion

This study analyzed healthcare claims and remits data to understand the trends and patterns of claim processing and payment. Descriptive statistics and data visualization techniques were used to gain insights into the data, while statistical analyses were performed to test hypotheses and determine the relationships between different variables. Predictive Analysis were performed to predict if the claims will be denied or paid, and the time taken to pay the claims.

Overall, this study provides useful insights into the processing and payment of healthcare claims and highlights the need for further research to better understand the factors affecting the time taken for claims reimbursement and the denial of claims by insurance companies. The findings of this study could also be useful for healthcare providers and insurance companies to improve their claim processing and payment systems.

Thank You

By:
Sravani Mahankali

Department:
Department Of Health Data Science, Saint Louis University - School of Medicine

Course:
HDS 5960 Health Data Science Capstone

Supervisor:
Sylvia Neil, Senior Director Business Intelligence

Instructors:
Divya S. Subramaniam, Ph.D., Mph, Paula Buchanan, Ph.D., Mph, Srikanth Mudigonda, Ph.D, Deepika Gopukumar, Ph.D