Nanyang Polytechnic

School of Information Technology

To: Dr Koh Noi Sian

Senior Lecturer

School of Information Technology

From: Yap Dong Sam (Team Leader)

210889C

Chee Bo Yu

214322H

Mark Alfonso

211684C

Tristan Hoh

212808Z

Subject: Milestone Report

Submission Date: 7th August 2022, Sunday 2359

**Contents**

1. Data Collection
2. Data Understanding
3. Data Preparation

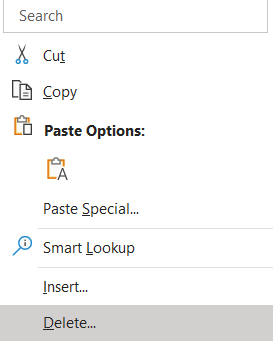
**1. Data Collection**

Finance (Bo Yu)

**Links:**

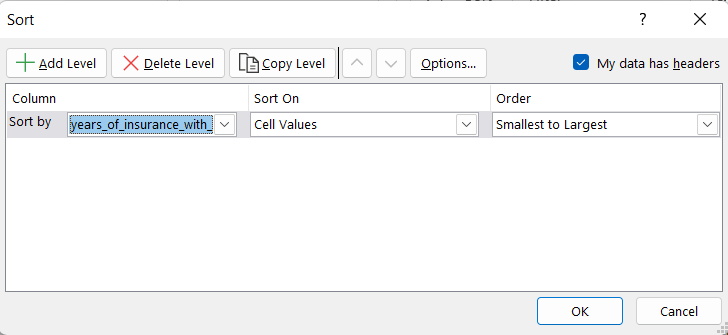
1. [Insurance (HealthCare)](https://www.kaggle.com/datasets/daminitiwari/insurance)
2. [Health Insurance Coverage](https://www.kaggle.com/datasets/hhs/health-insurance)
3. [Health Insurance data: Claims data](https://www.kaggle.com/datasets/ankush89/health-insurance-data-policy-and-claims-data?select=Health+Insurance+claims+data.csv)
4. [Hospital Charges for Inpatients](https://www.kaggle.com/datasets/speedoheck/inpatient-hospital-charges)
5. [Insurance](https://www.kaggle.com/datasets/gdeepakreddy/insurance)

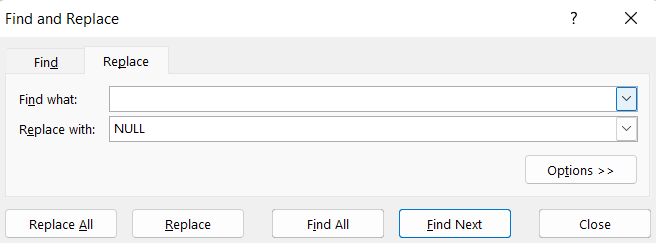
**Tools Used:**

* Excel

**Techniques Used:**

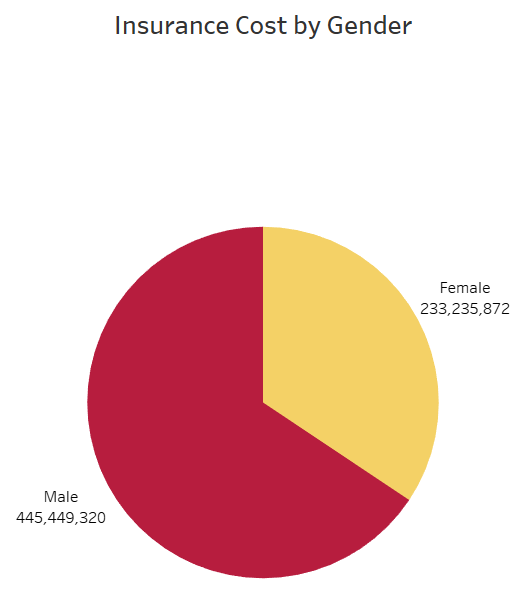
* Removing Unnecessary Columns(**Applicant\_ID, Cleansed\_inpatientCharges)** by using **Delete Columns**



* Sorting the column from smallest to largest (Years\_of\_Insurance) and Alphabet Order(**Type**,) by using **Sort & Filter** Function
* Adding NULL or 0 to blank values(**Year\_last\_admitted, State Medicaid Expansion (2016), Medicaid Enrollment (2013), Medicaid Enrollment Change (2013-2016), DOB, District, States, claim\_amt2, charges\_accommodation, charges\_nursing, charges\_package, charges\_consultant, charges\_surgery, charges\_pharmacy, charges\_ambulance, charges\_maternity, approved\_amt, deducted\_amt)** by using **Find and Replace** Function

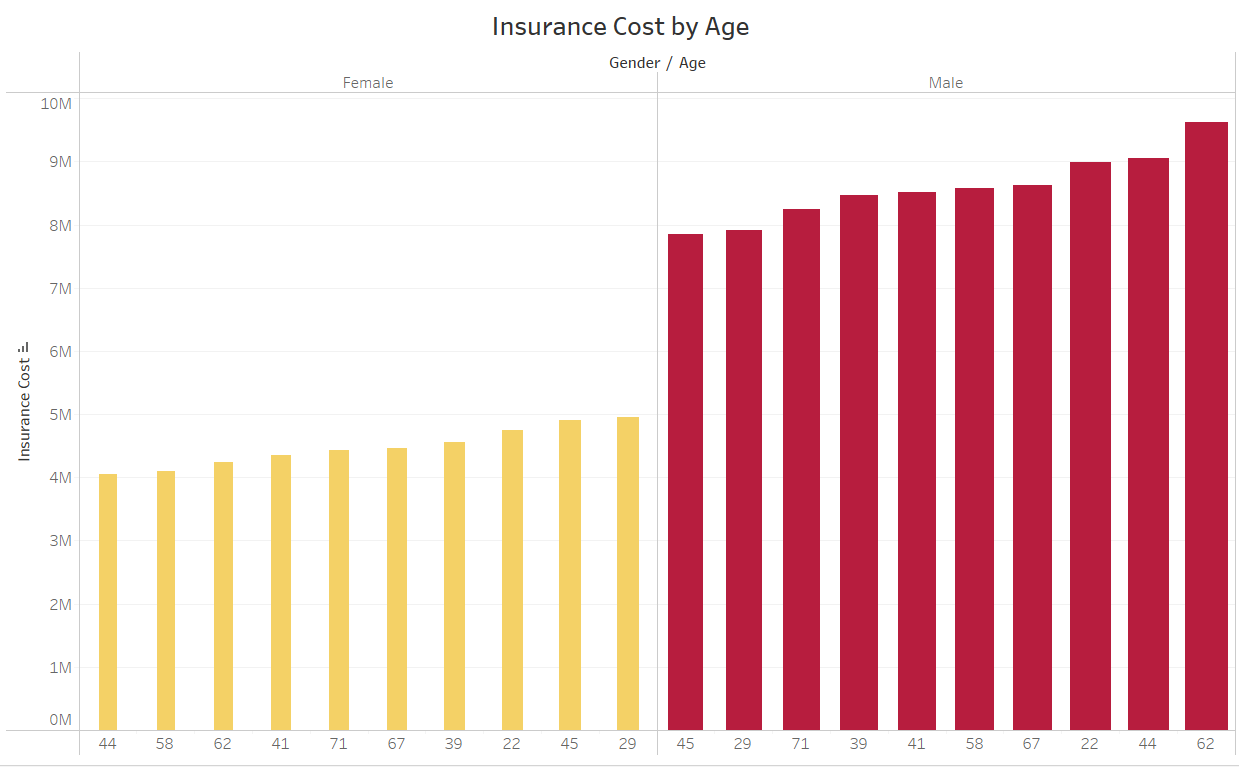
**2. Data Understanding**

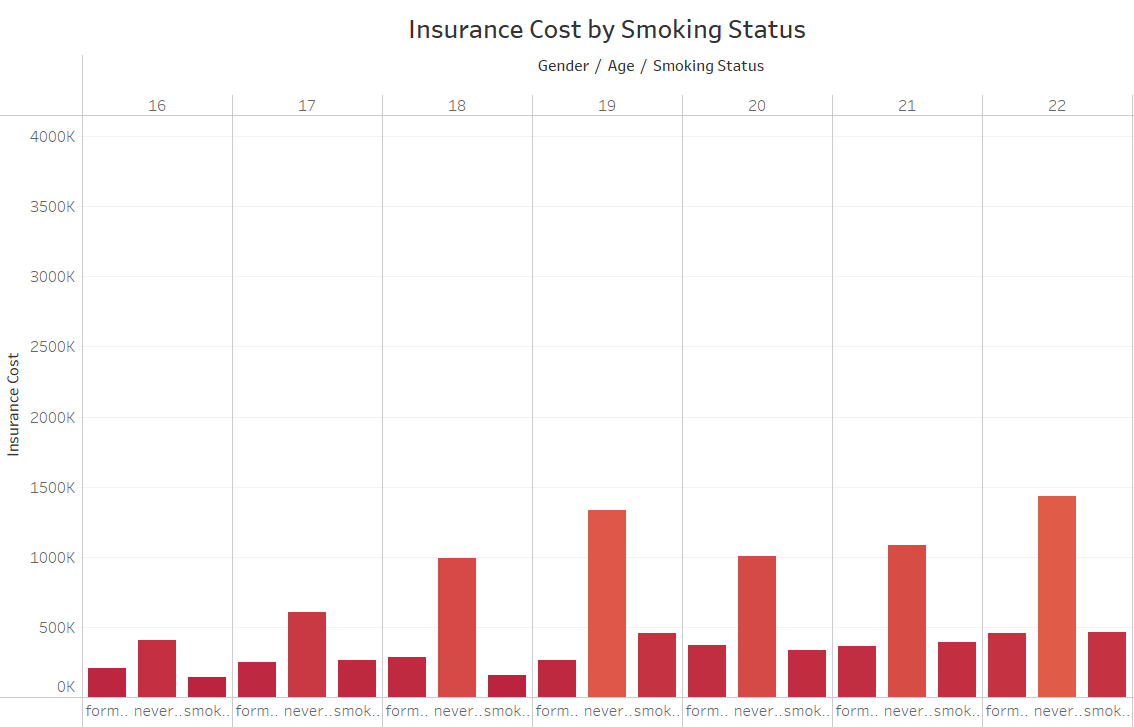
Finance (Bo Yu)



This chart shows between **Male** and **Female** Insurance Cost with **445,449,320** and **233,235,872** respectively and **Gender** and **Insurance Cost** details when hovering.

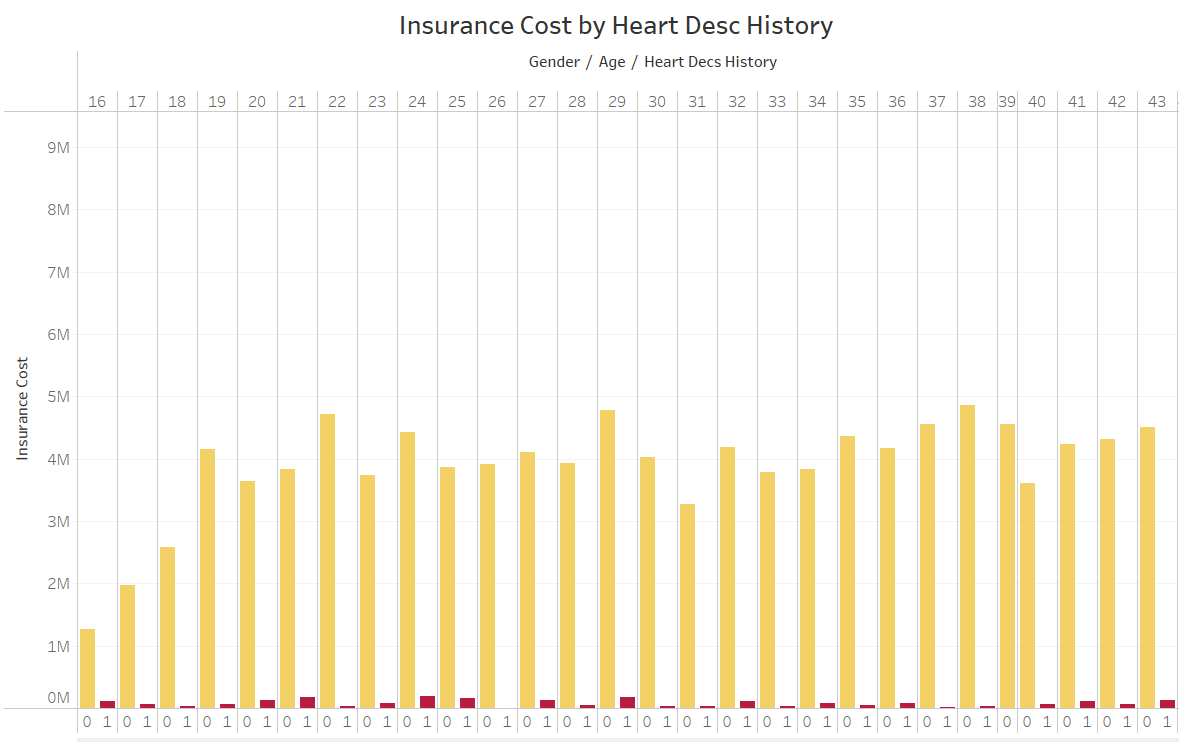
This chart also shows that there are more insurance costs toward male and it may be due to females getting more discounts on the plans where they need to give birth.

This chart shows the **Top 10** **Male** and **Female** Insurance Cost by Age. The top 1 for **male** is 62 and top 1 for **female** is 29. This chart also shows Male insurance is more expensive than Female when getting older and it may be due to early health conditions discovered when young.

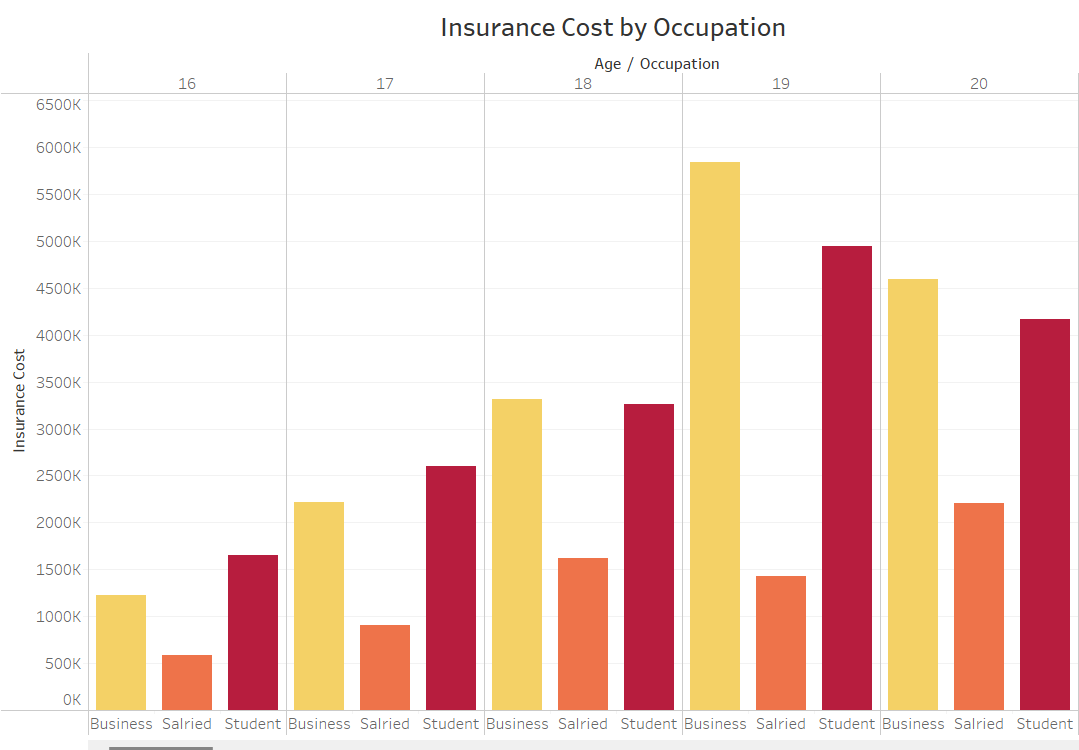


This chart shows between **Male** and **Female** Insurance Cost by Smoking Status from **age** 16 to 74 where it is classified by never smoked, formly smoked and smokes.

This chart also shows that people who don't smoke need more insurance coverage.

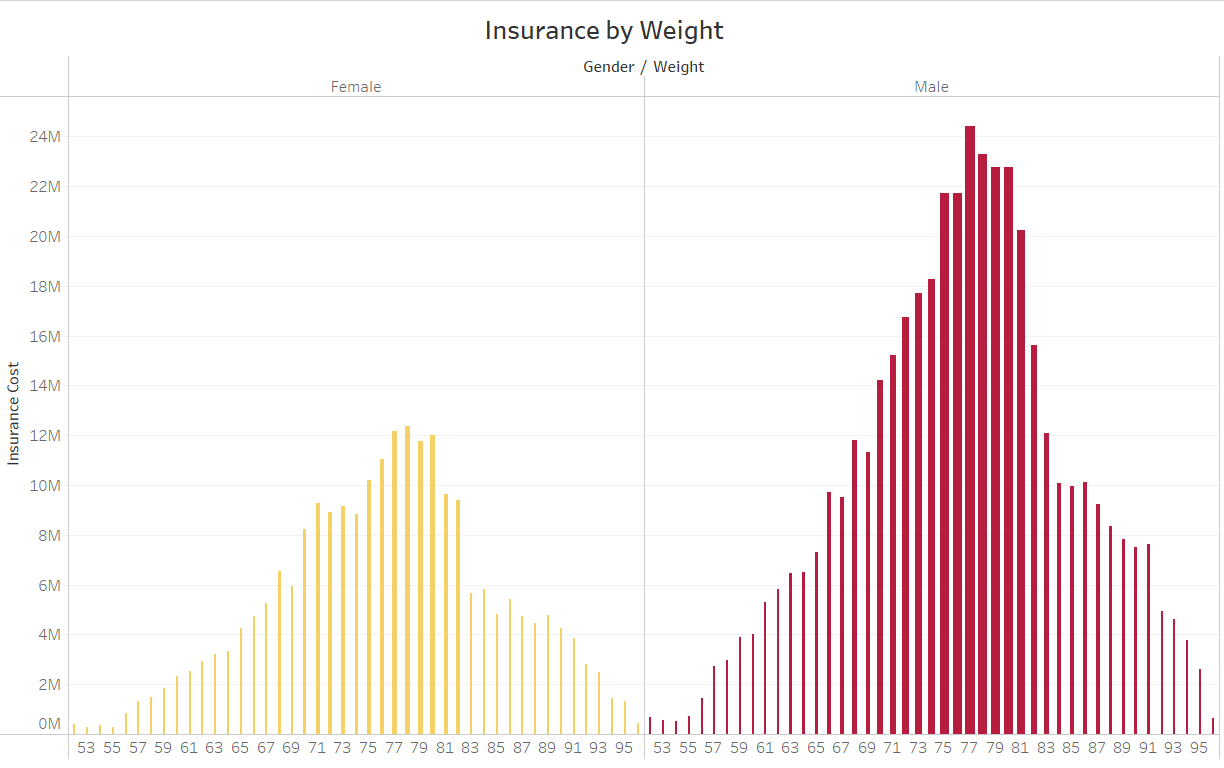
This chart shows the difference between **Male** and **Female** Insurance Cost by **Age** from 16 to 74 and **Heart Desc History** with **1** means have heart desc and **0** mean no heart desc.

This chart also shows that people who don't have Heart Desc History need to pay more money on insurance compared to those with Heart Desc History

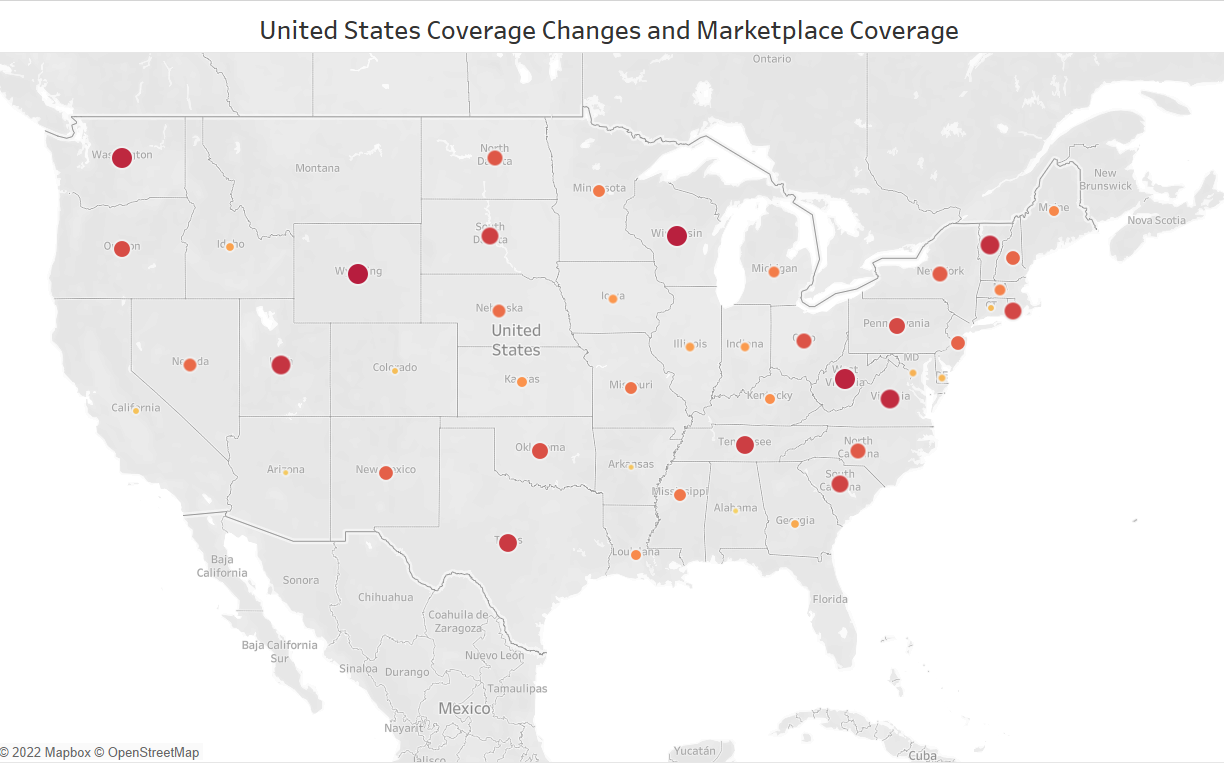


This chart shows Insurance Cost by **Age** from 16 to 74 and by **Occupation** which includes Business, Salried and Student.

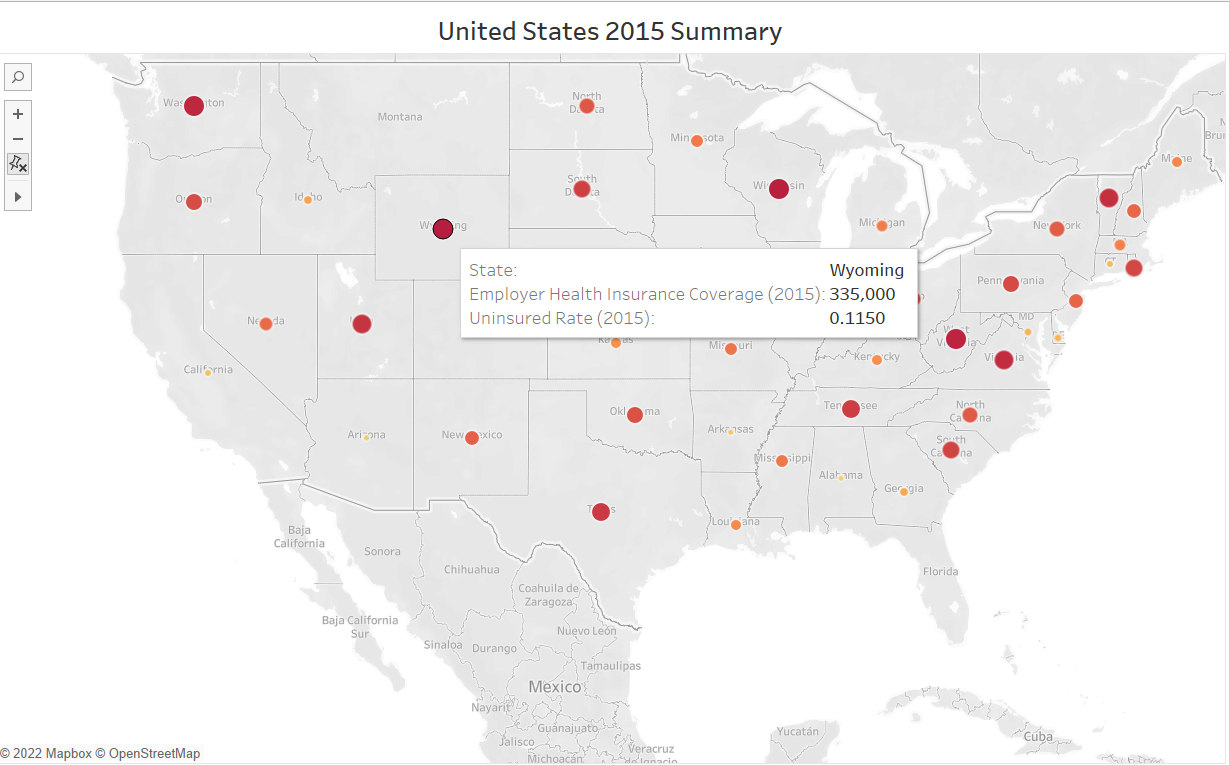
This chart also shows that people whose occupation is a Business or Student need more insurance when getting older and it may be due to their workplace asking them to buy healthcare insurance.

This chart shows Insurance by **Weight** where we can see an increasing trend from 53 to 81 but an unstable trend after 81 for females and an increasing trend from 53 to 77 but an unstable trend after 77 for male.

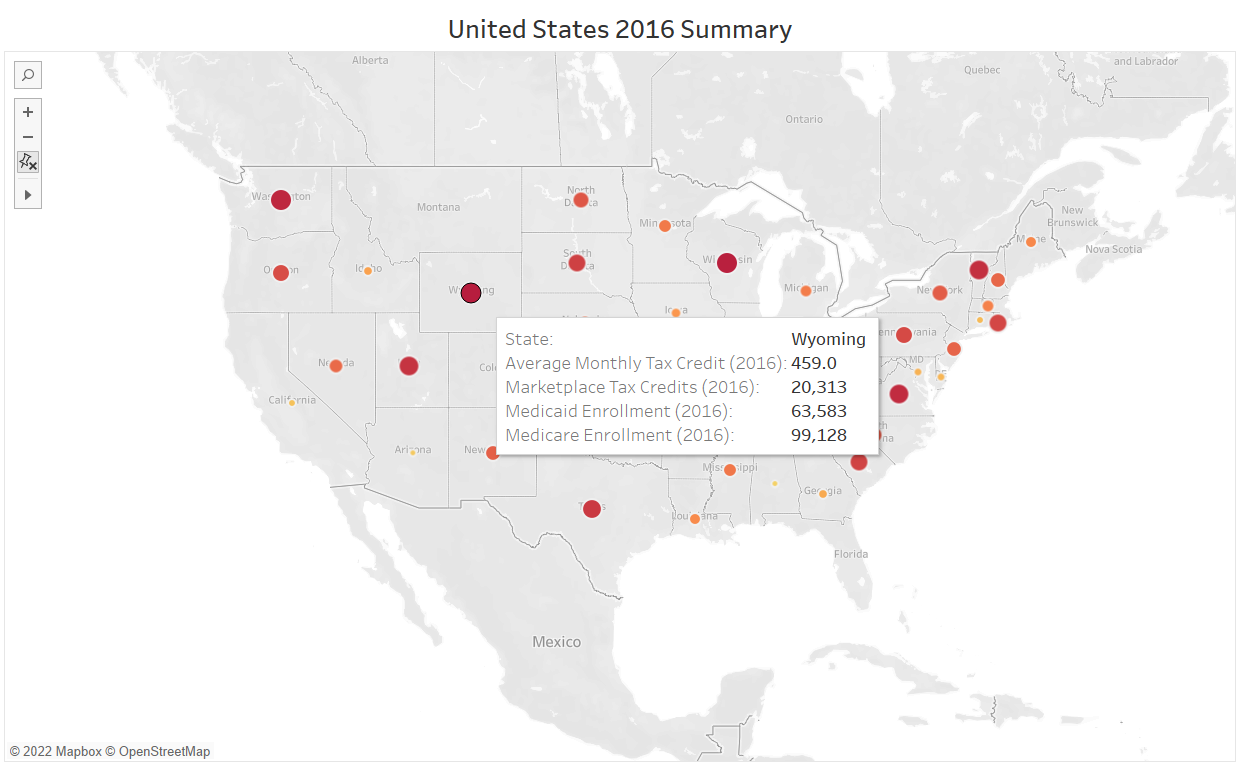
This chart also shows that people who are heavier need insurance which covers a lot of health conditions as shown on the graph.

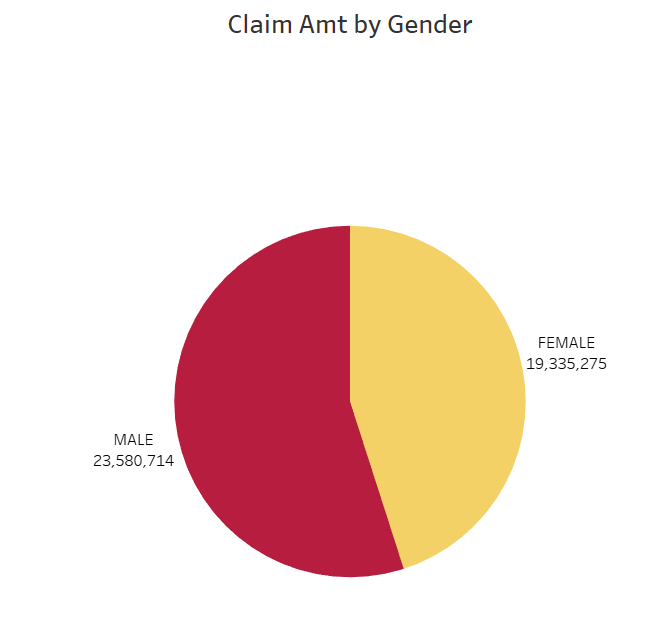


This chart shows **United States Coverage Changes and Marketplace Coverage with** Health Insurance Coverage Change (2010-2015), Medicaid Enrollment Change (2013-2016) and Uninsured Rate Change (2010-2015) details.



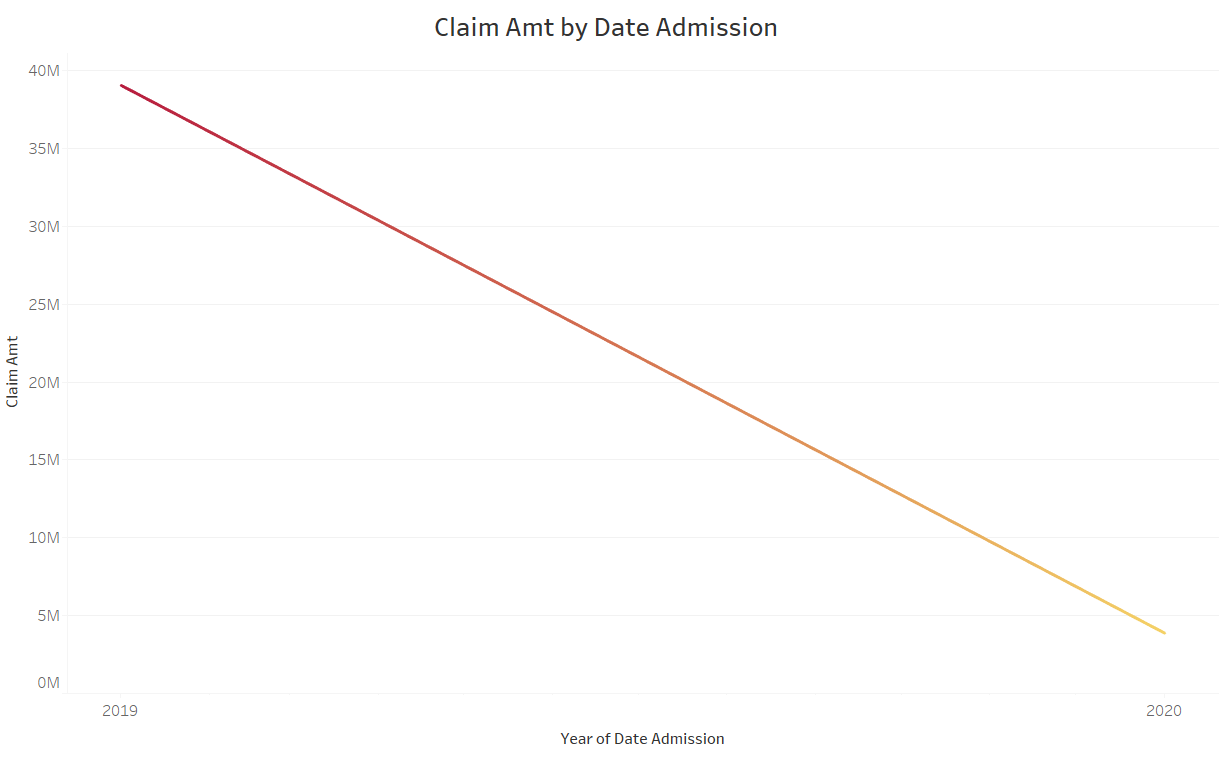
This chart shows the **United States 2015 Summary with** **Employer Health Insurance Coverage (2015), Uninsured Rate (2015)** and **State** details. We can see from the chart the highest coverage and rate is Wyoming with 335,000 and 0.1150 respectively

This chart shows the **United States 2016 Summary with** **Average Monthly Tax Credit (2016), Marketplace Tax Credits (2016), Medicaid Enrollment (2016)** and **Medicare Enrollment (2016)** details. We can see from the chart the highest average monthly tax credits, marketplace tax credits, medicaid enrollment and medicare enrollment is Wyoming with 459.0, 20,313, 63,583 and 99,128 respectively.

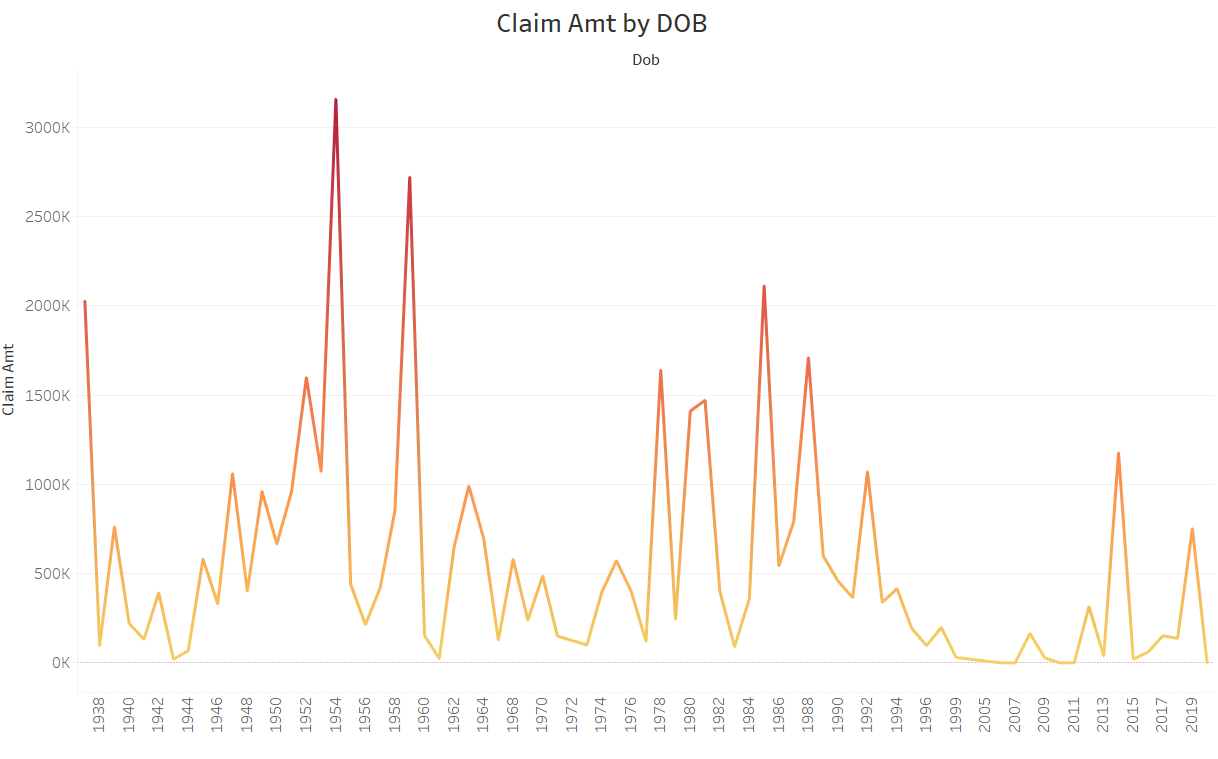


This chart shows the **Male** and **Female** Claim Amt with **23,580,714** and **19,335,275** respectively and **Gender** and **Claim Amt** details when hovering.

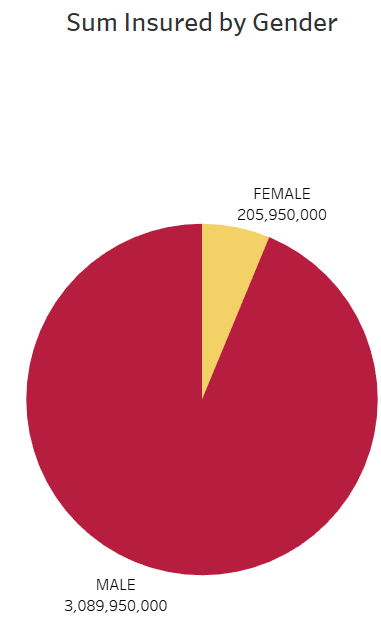
This chart also shows that there are more claims by Male than females. It may be due to more man elderly people who need the money to pay the hospital bills.



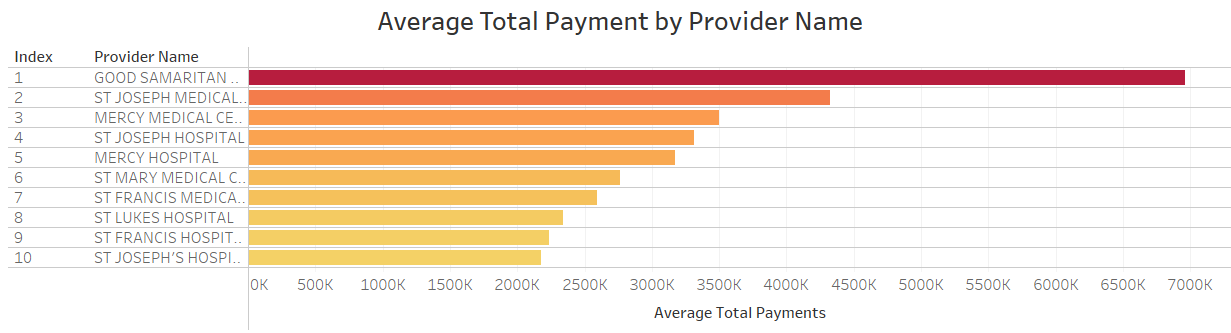
This chart shows Claim Amt by Date Admission from 2019 to 2020 with **3,871,013** and **39,044,976** respectively, it shows a **decreasing trend** from 2019 to 2020 and it may be due to Lockdown in 2020 which affects the Claim Amt.



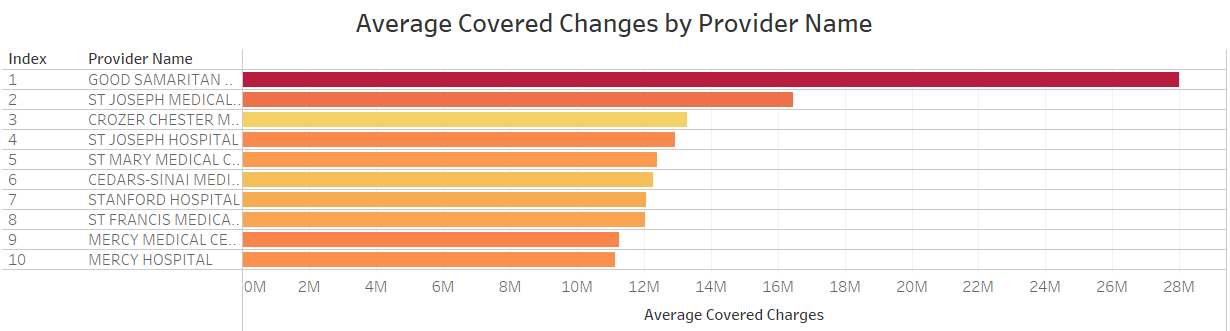
This chart shows Claim Amt by DOB from 1938 to 2019 with **2,026,480** and **751,788** respectively andit shows a **decreasing trend** from 1938 to 2019.



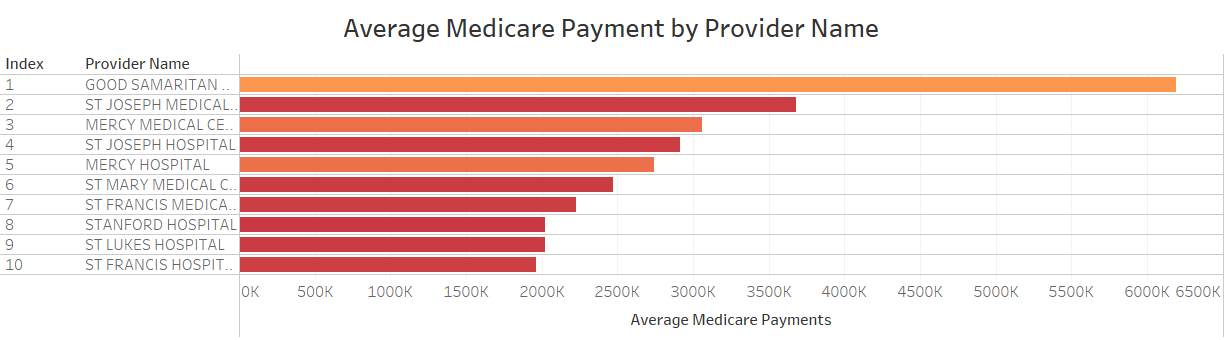
This chart shows the **Sum Insured by Gender** with Males having 3,089,950,000 sum insured and female having 205,950,000 sum insured. This chart also shows that the insurance company pays more money to Male than females if there are any **unpredictable events.**



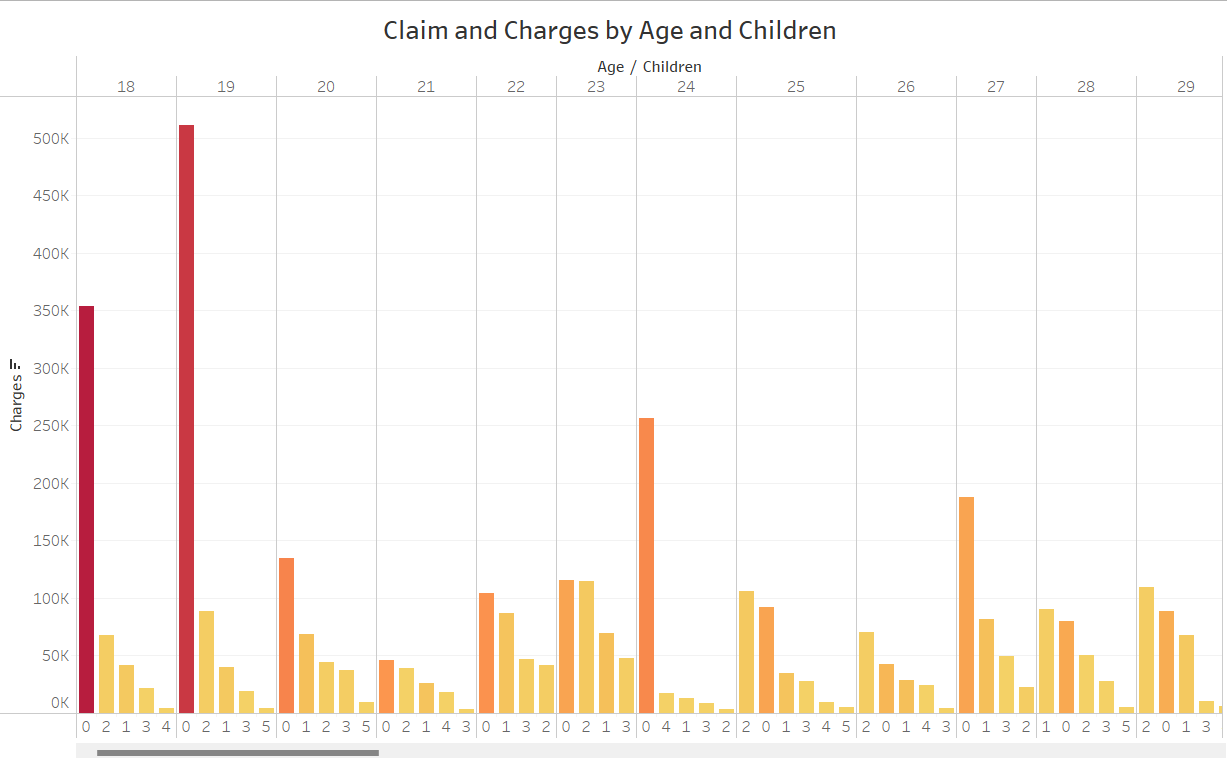
This chart shows the **Average Total Payment by Provider Name in the Top 10** with Good Samarian Hospital, ST Joesph Medical Center and Mercy Medical Center ranking in the top 3.



This chart shows the **Average Covered Changes by Provider Name in Top 10**  with Good Samarian Hospital, ST Joesph Medical Center and Crozer Chester Medical Center ranking in the top 3.

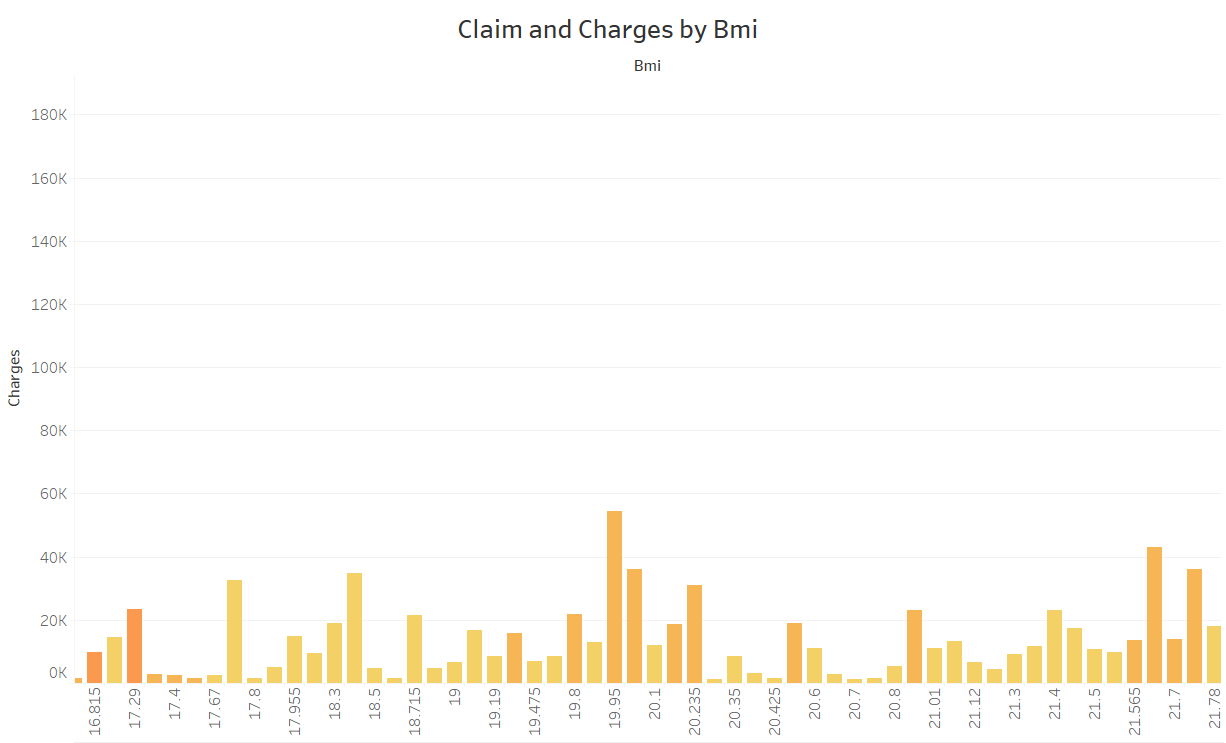


This chart shows the **Average Medicare Payment by Provider Name in the Top 10**  with Good Samarian Hospital, ST Joesph Medical Center and Mercy Medical Center ranking in the top 3.



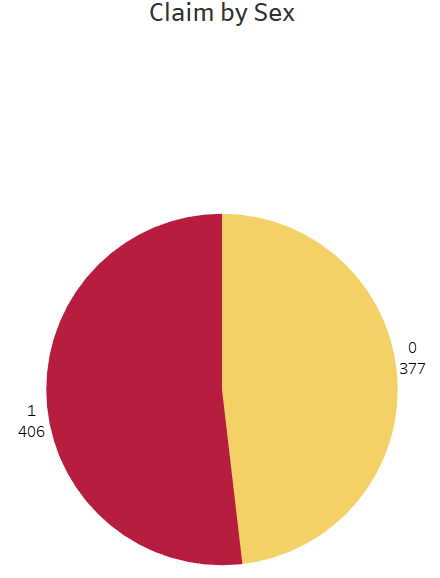
This chart shows **Charges by Age (18 to 64) and Children (0 to 5)** and when you hover on the chart we will be able to see the charges and insurance claim details.

This chart also shows when there are more children, there are lesser claims and charges compared to zero children.



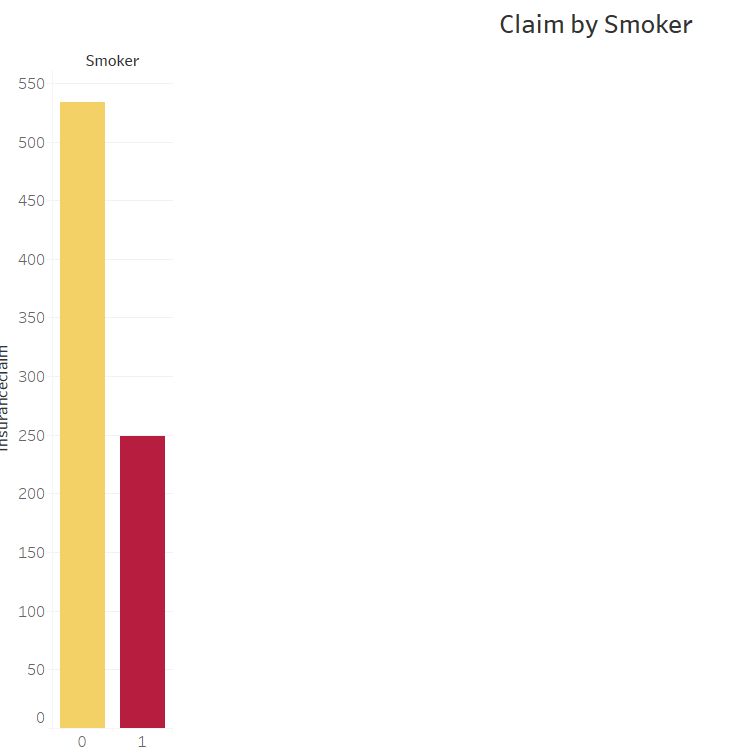
This chart shows **Charges by Bmi from 15 to 53** and when you hover on the chart we will be able to see the charges and insurance claim details.

This chart also shows that when the bmi increases , the claim and charges also increase and it may due to people with high bmi often go for health checkup



This chart shows the **Claim by Sex** where 1 mean have and 0 mean don’t have and when hover on the chart we will be able to see the charges and insurance claim details.

This chart also show that those will sex need to go for checkup for personal reason compared than those without which lead to a more result on the 1 than 0.



This chart shows **Claim by Smoker** where 1 means have smoke and 0 means no smoke and when you hover on the chart we will be able to see the charges and insurance claim details. It may be due to unbalanced data on the people who smoke.

**3. Data Preparation**

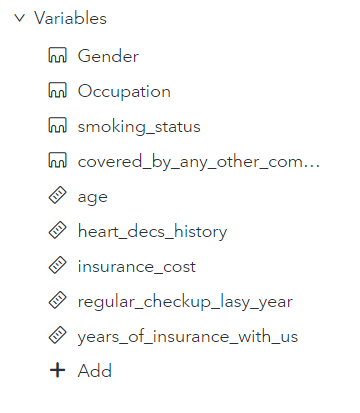
Finance (Bo Yu)

I plan to use 2 modeling techniques which consists of **Clustering** and **Linear Regression with Correlation Matrix** for my objectives.

For **Clustering**, we can use the **Cluster Diagram window** shows prediction ellipses and whether they are filled on each scatter plot for each Cluster ID, use **Parallel Coordinates Plot** to compare many quantitative variables together looking for patterns and relationships between themand use **Auto Chart** to see which Cluster Id has the largest amount to explore the availability of various subsidies such as insurance, how many are covered by them and to determine the crucial elements that contribute to good insurance.

For **Correlation Matrix,** we can use **Correlation of selected measures** to see which inputs will affect the target for example if the target is highly correlated with the inputs , we can use in the finding and if the inputs is has a high correlation between inputs , we cannot use it as it will affect the output of the findings.

For Linear Regression, we can **Fit Summary** to see which inputs with the values more 0.05 that affect the output of the findings to identify factors such as Uninsured Rate Change and enrollment that can reduce the cost of hospital bills and make sure that more than 50% of the patients in US are fully covered by insurance when paying hospital bills.



Chart, bar chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generatedChart, bar chart

Description automatically generated

A picture containing graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated