An Analysis on Covid-19: Gender, Age, and Health

Who is most likely to be hospitalized?

KU Analytics BootCamp: J. Freeman, J. Forbis, N. Tarver, K. St. Germain, S. Sweeny

Outline

- Abstract
- Questions/Hypothesis
- Data Cleanup & Exploration
- Data Analysis
  - Line Chart
  - Hypothesis 1: Gender vs. Hospitalizations
  - Hypothesis 2: Age vs. Hospitalizations
  - Hypothesis 3: Pre-Existing Conditions vs. Hospitalizations
- Summary
- Post Mortem
- Q&A
- Resources

Abstract

Background:

- On December 31st, 2019, China notified the World Health Organization (WHO) about a cluster of pneumonia cases in Wuhan. In a few weeks, WHO identified a novel coronavirus as the cause of this cluster.[1]

- By the end of January of 2020, the United States reported its first case. In just under two months, the U.S. had exceeded 100,000 cases, becoming the new epicenter of the pandemic.[1]

- Covid-19's high infection rate has led to high hospital capacity and bed use.

Core Question:

- Who is most likely to be hospitalized?

1. *Archived: WHO Timeline - COVID-19*. (2020, April 28). World Health Organization. https://www.who.int/news/item/27-04-2020-who-timeline---covid-19

Questions

- **Core Question:** Who is most likely to be hospitalized?
  - **Hypothesis 1:** If males are more susceptible to Covid-19, then more hospitalizations will occur among male patients.
  - **Hypothesis 2:** If the elderly are more susceptible to Covid-19, then more hospitalizations will occur among elderly patients.
  - **Hypothesis 3:** If a person has has a pre-existing medical condition and contracts COVID-19, then they are more at risk to be hospitalized.

Data

- For our hypothesis hospitalization data was found for various demographic groups across the United States. Initial desired data needed to contain information relating age, ethnicity, and gender to hospitalizations in order to support the hypotheses.
- In order to accomplish this various sources of data were looked at on COVID-19 available to the public like from the CDC or the Atlantic's COVID Tracking project.
- In the end we chose the CDC's case surveillance data which provided us with over 8,000,000 cases with the age, ethnicity, and gender, as well as the hospitalization status .
- This dataset also provided us with whether or not the patient was put in the ICU or had a pre-existing condition medical condition.

Data Wrangling & Tidying Step 1: Importing Data into Pandas Dataframe

The first step of the data wrangling and tidying process is to import the necessary modules, store the path of the file into a python variable, and then use that variable with pandas methods to display the file in a data frame format.

The data frame should always be inspected before going to the next step. The .head() method can be used to inspect a clip of the entire data frame to prevent the computer from loading more data than necessarily needed at the time.

```
# Importing Relevant Dependencies
import pandas as pd
from matplotlib import pyplot as plt
import scipy.stats as st
import datetime as dt
import numpy as np


cdcPath = "../../../COVID-19_Case_Surveillance_Public_Use_Data.csv"
```

```
# Create data frame from CSV file variable(path) using pandas .read_csv() method
cdcDf = pd.read_csv(cdcPath, low_memory=False)

# Print out first 10 rows for inspection
cdcDf.head(3)
```

	cdc_report_dt	pos_spec_dt	onset_dt	current_status	sex	age_group	Race and ethnicity (combined)	hosp_yn	icu_yn	death_yn	medcond_yn
0	2020/11/10	2020/11/10	NaN	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	Unknown	No	No
1	2020/11/14	2020/11/10	2020/11/10	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No	No	No
2	2020/11/19	2020/11/10	2020/11/09	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No	No	No

Data Wrangling & Tidying Step 2: Remove Unnecessary Columns

After first inspection, it is evident there are unnecessary columns pertaining to the scope of the hypothesis. These columns should be removed before moving forward for readability and analysis purposes.

In this case only the Report Date, Current Status, Sex, Age Group, Race/Ethnicity, Hospitalization Status, and Pre-Existing Medical Condition columns were chose to be kept for further analysis.

```
# Dropping unnecessary columns
cutCdcDf = cdcDf.drop(['pos_spec_dt', 'onset_dt','icu_yn','death_yn'], axis=1)

cutCdcDf.head()
```

	cdc_report_dt	current_status	sex	age_group	Race and ethnicity (combined)	hosp_yn	medcond_yn
0	2020/11/10	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No
1	2020/11/14	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No
2	2020/11/19	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No
3	2020/11/14	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	Missing	Missing
4	2020/11/13	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	Yes

Data Wrangling & Tidying Step 3: Removal of Incomplete or Inaccurate Data

Many times data is given with incomplete, duplicate, or inaccurate data. This data should often be removed. To identify this data the .value_counts() method can be used to return all the different values found in a column and a count of each.

After identifying the menacing data, they can easily be removed using pandas methods in conjunction with conditional arguments.

An example of this process can be seen in the figure to the right.

```
# Checking values in the Age Groups
ethUnkDf['age_group'].value_counts()
```

```
20 - 29 Years    951360
30 - 39 Years    799407
40 - 49 Years    742803
50 - 59 Years    727465
60 - 69 Years    531308
10 - 19 Years    503431
70 - 79 Years    316264
80+ Years        254563
0 - 9 Years      185198
Unknown           10507
Name: age_group, dtype: int64
```

```
# Dropping unknown values from the age groups
unkAgeDf = ethUnkDf[ethUnkDf['age_group'] != 'Unknown']

# Checking for accuracy
unkAgeDf.head()
```

	cdc_report_dt	current_status	sex	age_group	Race and ethnicity (combined)	hosp_yn	medcond_yn
0	2020/11/10	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No
1	2020/11/14	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No
2	2020/11/19	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No
3	2020/11/14	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	Missing	Missing

Data Wrangling & Tidying Step 4: Final Tidying of Desired Dataframe

Once only the complete, accurate data is contained in the data frame it is often helpful to sort the data in some sort of way. The dataframe in interest was best sorted by time so that it could be analyzed on a time basis if so chosen.

Unfortunately pandas does not automatically reindex the data so it is often also desirable to do so. This is done with the .reset_index method().

Lastly, the data is almost always given with column headers that contain underscores and names that do not clearly explain what it is that the column values contain. To help with this, the .rename() method can be used to give the column headers clean, descriptive titles.

An image of the data frame in which is sorted by time, reindexed, and has renamed columns, can be seen to the right.

```
# Renaming Column Headers for Visualization and Analyis Purposes
rnCdcDf = cleanCdcDf.rename(columns={"cdc_report_dt": "CDC Report Date",
                                     "current_status": "Covid Status",
                                     "sex": "Gender",
                                     "age_group": "Age Group",
                                     "Race and ethnicity (combined)": "Race/Ethnicity",
                                     "hosp_yn": "Hospitalized",
                                     "medcond_yn": "Pre-Existing Condition"})
# Sort values by CDC Report Date
sorted_CdcDf = rnCdcDf.sort_values("CDC Report Date", ascending=True)

# Reset index and save to new variable
resetCdcDf = sorted_CdcDf.reset_index(drop=True)

resetCdcDf.head()
```

	CDC Report Date	Covid Status	Gender	Age Group	Race/Ethnicity	Hospitalized	Pre-Existing Condition
0	2020/01/01	Laboratory-confirmed case	Female	60 - 69 Years	White, Non-Hispanic	No	Yes
1	2020/01/01	Laboratory-confirmed case	Male	40 - 49 Years	Black, Non-Hispanic	Yes	No
2	2020/01/02	Laboratory-confirmed case	Male	30 - 39 Years	White, Non-Hispanic	No	Yes
3	2020/01/09	Laboratory-confirmed case	Female	40 - 49 Years	White, Non-Hispanic	No	No
4	2020/01/11	Laboratory-confirmed case	Male	60 - 69 Years	White, Non-Hispanic	No	Yes

Data Wrangling & Tidying Step 5: Create specific data frames for each Hypothesis.

To better focus on the data the are in the scope of each hypothesis specifically, it is helpful to create new data frames that only contain data pertaining to each hypothesis.

Clips of the data frames specific to Gender vs. Hospitalization, Age Group vs. Hospitalization, and Pre-Existing Condition vs. Hospitalization can be seen below.

```
## GENDER vs. HOSPITALIZATION

# Data Frame Specific to the Analysis of Gender
patient_sex_df = resetCdcDf[["CDC Report Month",
patient_sex_df.head()
```

	CDC Report Month	Gender	Hospitalized
0	January	Female	No
1	January	Male	Yes
2	January	Male	No
3	January	Female	No
4	January	Male	No

```
## AGE GROUP vs. HOSPITALIZATION

# Data Frame Specific to the Analysis of Age Group
# Removed uneeded column
patient_age_df = resetCdcDf[["CDC Report Month", "
patient_age_df.head()
```

	CDC Report Month	Hospitalized	Age Group
0	January	No	60 - 69 Years
1	January	Yes	40 - 49 Years
2	January	No	30 - 39 Years
3	January	No	40 - 49 Years
4	January	No	60 - 69 Years

```
## PRE-EXISTING CONDITION vs. HOSPITALIZATION

# Data Frame Specific to the Analysis of Pre-Ex
patient_premed_df = resetCdcDf[["CDC Report Mon
patient_premed_df.head()
```

	CDC Report Month	Hospitalized	Pre-Existing Condition
0	January	No	Yes
1	January	Yes	No
2	January	No	Yes
3	January	No	No
4	January	No	Yes

Data Wrangling & Tidying Step 6: Line Chart Prep Process

The data imported from the CDC contained values containing only strings and boolean values. Due to this some reorganization and manipulation of the data was necessary in order to analyze the data only a time-varying basis.

The first issue that occured in this process came from the 'CDC Report Date' column. This column contained the date, but in a string format. To overcome this issue Pandas pd.to_datetime() method was used to change the string type to a date type.

The next speed bump was grouping the data by month, and labeling the row accordingly. This was done by changing the data to a single integer representing each month and then binning the data by this integer and returning a string of the month in a new column..

This process from start to finish as well as the clip of the new columns produced from binning can be see in the figure below.

```
# Convert Date column to correct "date" type
resetCdcDf['CDC Report Date'] = pd.to_datetime(resetCdcDf['CDC Report Date'], format= '%Y/%m/%d')

# Create new column for months of the date for grouping purposes and change to "numerical" type
resetCdcDf['CDC Report Month (#)'] = resetCdcDf['CDC Report Date'].dt.strftime('%m')
resetCdcDf['CDC Report Month (#)'] = pd.to_numeric(resetCdcDf['CDC Report Month (#)'])

# Create bins to aggregate the data into months and use pd.cut() method to bin the data into a new column
bins = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]
months = ["January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "November"]

resetCdcDf['CDC Report Month'] = pd.cut(resetCdcDf['CDC Report Month (#)'], bins, labels=months, include_lowest=False)
resetCdcDf.head()
```

	CDC Report Date	Covid Status	Gender	Age Group	Race/Ethnicity	Hospitalized	Pre-Existing Condition	CDC Report Month (#)	CDC Report Month
0	2020-01-01	Laboratory-confirmed case	Female	60 - 69 Years	White, Non-Hispanic	No	Yes	1	January
1	2020-01-01	Laboratory-confirmed case	Male	40 - 49 Years	Black, Non-Hispanic	Yes	No	1	January
2	2020-01-02	Laboratory-confirmed case	Male	30 - 39 Years	White, Non-Hispanic	No	Yes	1	January
3	2020-01-09	Laboratory-confirmed case	Female	40 - 49 Years	White, Non-Hispanic	No	No	1	January
4	2020-01-11	Laboratory-confirmed case	Male	60 - 69 Years	White, Non-Hispanic	No	Yes	1	January

Data Analysis

Hypothesis 1

If males are more susceptible to Covid-19, then more hospitalizations will occur among male patients.

- To answer the hypothesis above, the dataset, "cleanedCaseSurveillance.csv" was used to analyze the population of inpatient males and females.

Step 1:

```
# Creating Filepath
filepath = '../Resources/cleanedCaseSurveillance.csv'

# Reading CSV File
df = pd.read_csv(filepath)

# Printing DataFrame
df.head()
```

	CDC Report Date	Covid Status	Gender	Age Group	Race/Ethnicity	Hospitalized	Pre-Existing Condition
0	2020/11/10	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No
1	2020/11/14	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No
2	2020/11/19	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No
3	2020/11/13	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	Yes
4	2020/11/09	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No

Hypothesis 1

If males are more susceptible to Covid-19, then more hospitalizations will occur among male patients.

Step 2:

```
df = df[['Covid Status', 'Gender', "Hospitalized"]]

df.head(10)
```

	Covid Status	Gender	Hospitalized
0	Laboratory-confirmed case	Male	No
1	Laboratory-confirmed case	Male	No
2	Laboratory-confirmed case	Male	No
3	Laboratory-confirmed case	Male	No
4	Laboratory-confirmed case	Male	No
5	Laboratory-confirmed case	Male	No
6	Laboratory-confirmed case	Male	No
7	Laboratory-confirmed case	Male	No
8	Laboratory-confirmed case	Male	No
9	Laboratory-confirmed case	Male	No

Hypothesis 1

If males are more susceptible to Covid-19, then more hospitalizations will occur among male patients.

Step 3:

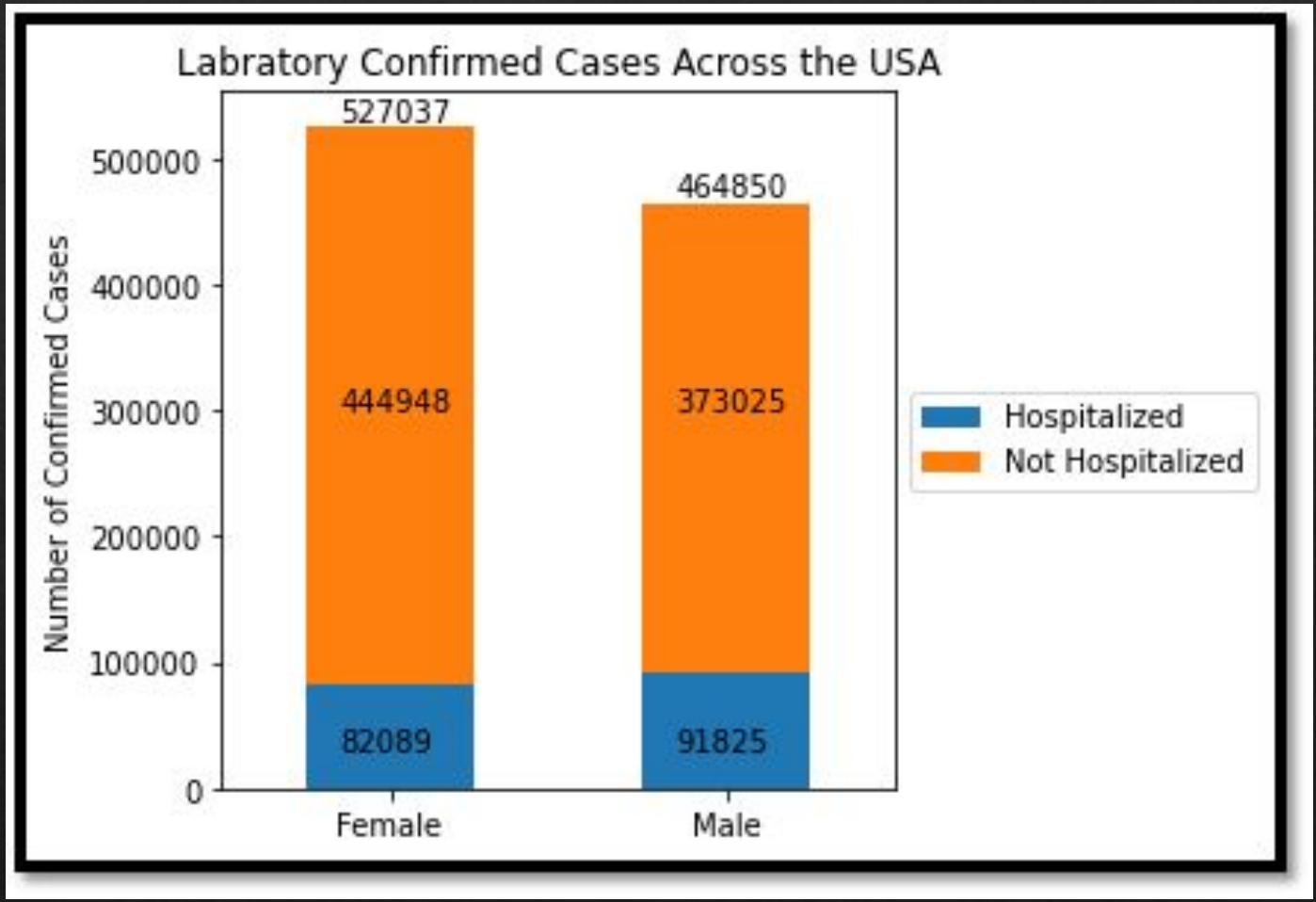
```
# Creating DataFrame of Hospitalized Males and Females
hosp_male_df = df.loc[(df['Gender'] == 'Male') & (df['Hospitalized'] == 'Yes')]
hosp_female_df = df.loc[(df['Gender'] == 'Female') & (df['Hospitalized'] == 'Yes')]

hosp_male_size = len(df.loc[(df['Gender'] == 'Male') & (df['Hospitalized'] == 'Yes')])
hosp_female_size = len(df.loc[(df['Gender'] == 'Female') & (df['Hospitalized'] == 'Yes')])

no_hosp_male_size = len(df.loc[(df['Gender'] == 'Male') & (df['Hospitalized'] == 'No')])
no_hosp_female_size = len(df.loc[(df['Gender'] == 'Female') & (df['Hospitalized'] == 'No')])
```

Hypothesis 1

If males are more susceptible to Covid-19, then more hospitalizations will occur among male patients.

Discussion:

- Women have the most Laboratory Confirmed Cases: 527,037
- Men are hospitalized more than women, but only by 5.6%.
- Data in this dataset is categorical, a statistical test cannot be used to accept or reject the hypothesis.
- The graph suggests that women are more susceptible to Covid-19, but men are hospitalized more.
- The hypothesis cannot be accepted or reject, further analysis is needed.



Hypothesis 2

If the elderly are more susceptible to Covid-19, then more hospitalizations will occur among elderly patients.

The group chose to analyze the the age groups by two separate divisions, above 50 years old, and below. To do this the data somehow had to be divided from their original grouping labels to another. This was tricky bc it could be done several ways.

To do this a copy of the cleaned data was made and the .replace() method was used on this copy to divide the age groups into two divisions, “Older” and “Younger”, to reference above or below 50 years of age.

It was chosen to be done this way because when trying to use many methods such as .replace() on a slice of a dataframe, pandas will throw index error messages. Also, it is safe to make changes like this on a copy so the original data frame can also be later referenced.

A new sliced data frame from this copy from explanation above can be seen to the right.

```
# Now, Make a Data Frame Specific to the
patient_age_df = renamedCdcDf1[["CDC Repo
patient_age_df.head(10)
```

	CDC Report Month	Hospitalized	Age Division
0	January	No	Older
1	January	Yes	Younger
2	January	No	Younger
3	January	No	Younger
4	January	No	Older
5	January	Yes	Younger
6	January	Yes	Older
7	January	No	Older
8	January	No	Older
9	January	No	Older

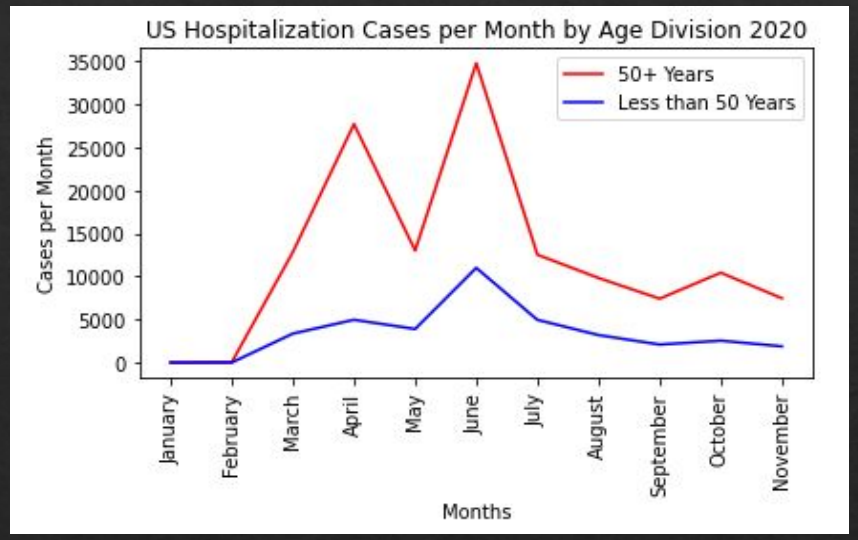
Hypothesis 2

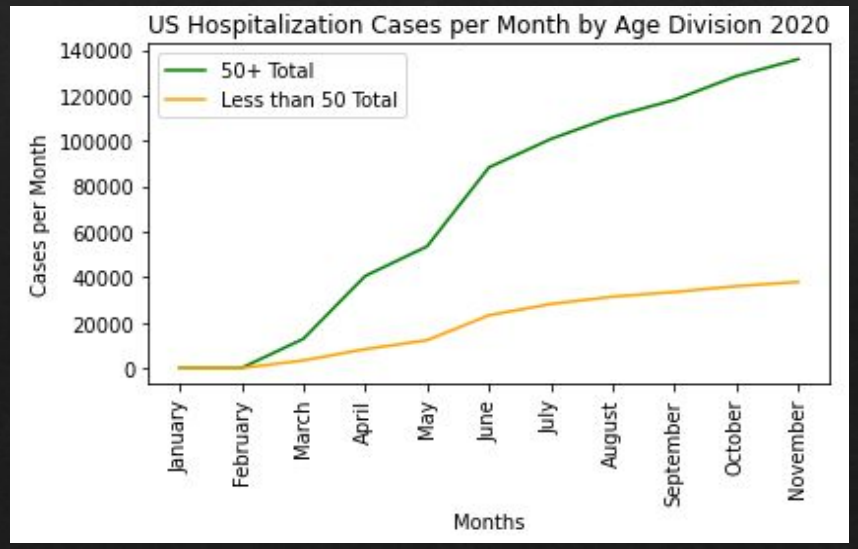
If the elderly are more susceptible to Covid-19, then more hospitalizations will occur among elderly patients.

After dividing the age groups into two divisions the number of cases of patients who weren't hospitalized were filter out of the data and the remainder were plotted of the course of the yearly simply to display a comparison between two.

The line charts to the right display the data in a unique way. The chart on the top right displays the monthly average for each division of the year while the figure in the bottom right displays the running total over the course of the year.

During the initial analysis of the generated line plots it can be noticed immediately that persons over the age of 50 are much more susceptible to COVID. The chart above shows the rate of change of the hospitalizations over the course of the year while the chart below shows the sum of the hospitalizations over the year, both of which support the hypothesis directly.



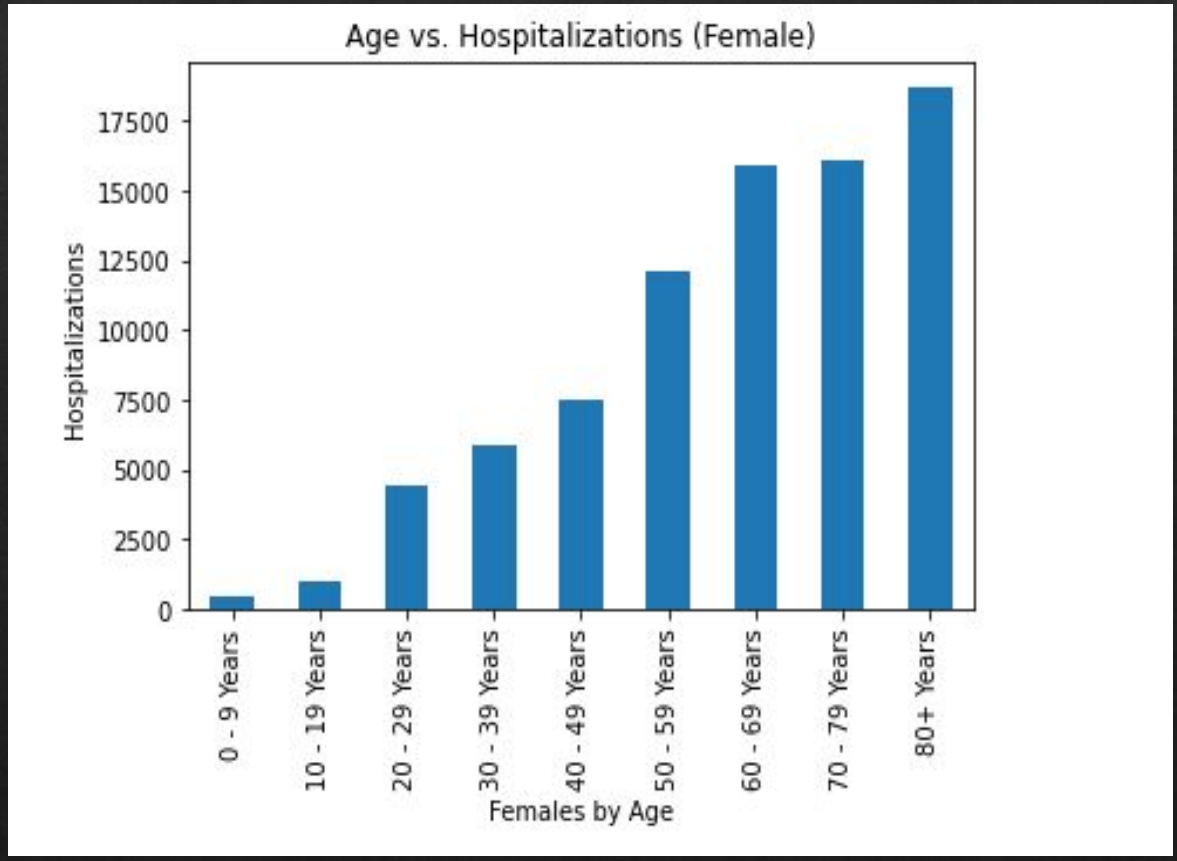


Hypothesis 2

If the elderly are more susceptible to Covid-19, then more hospitalizations will occur among elderly patients.

Discussion:

- Women in their 80's+ are the most hospitalized age group
- Women who are 60-69 year olds & 70-79 year olds are the 2nd largest age groups hospitalized
- Women in their 50's are the 3rd age group to be hospitalized
- This information proves that elderly women are more hospitalized and therefore more susceptible to COVID

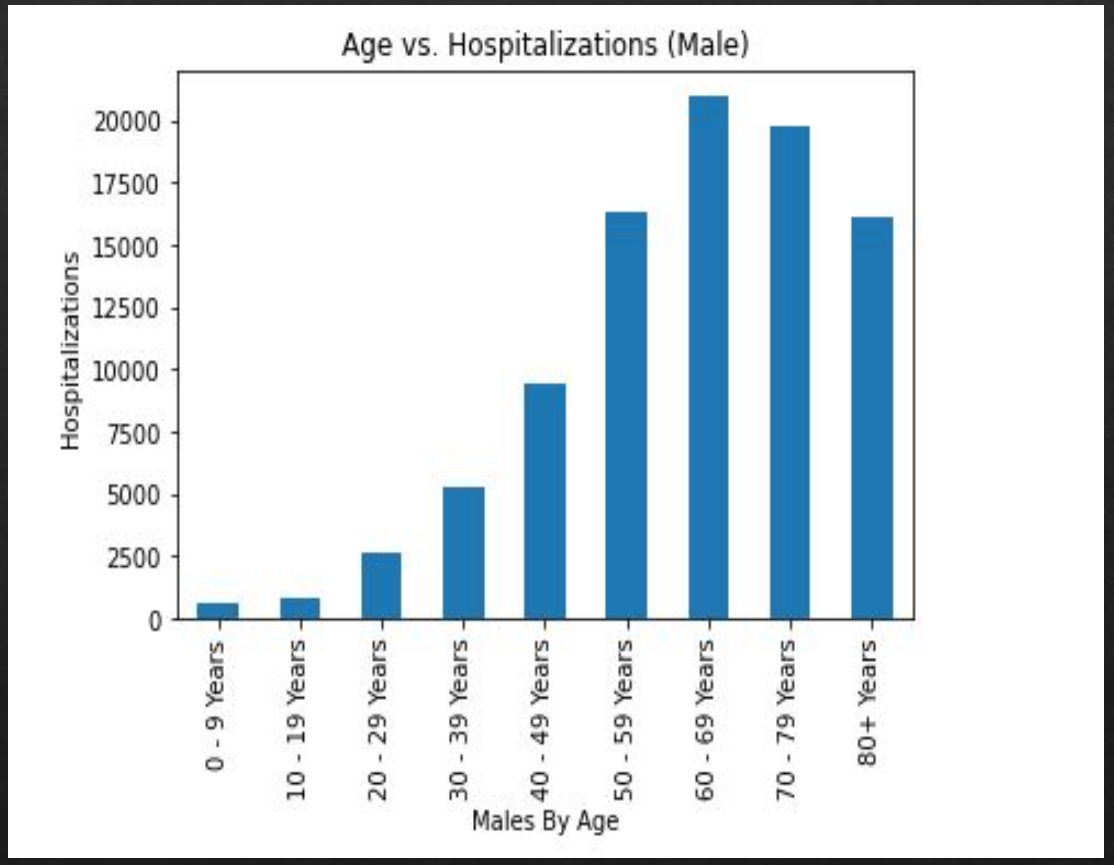


Hypothesis 2

If the elderly are more susceptible to Covid-19, then more hospitalizations will occur among elderly patients.

Discussion:

- The age group of Men with the most hospitalizations is 60-69 year olds
- Men in who are 70-79 years old are the 2nd largest group to be hospitalized
- Men in their 50's & 80's are the 3rd largest age group to be hospitalized
- This information further supports our hypothesis that older men are susceptible to COVID

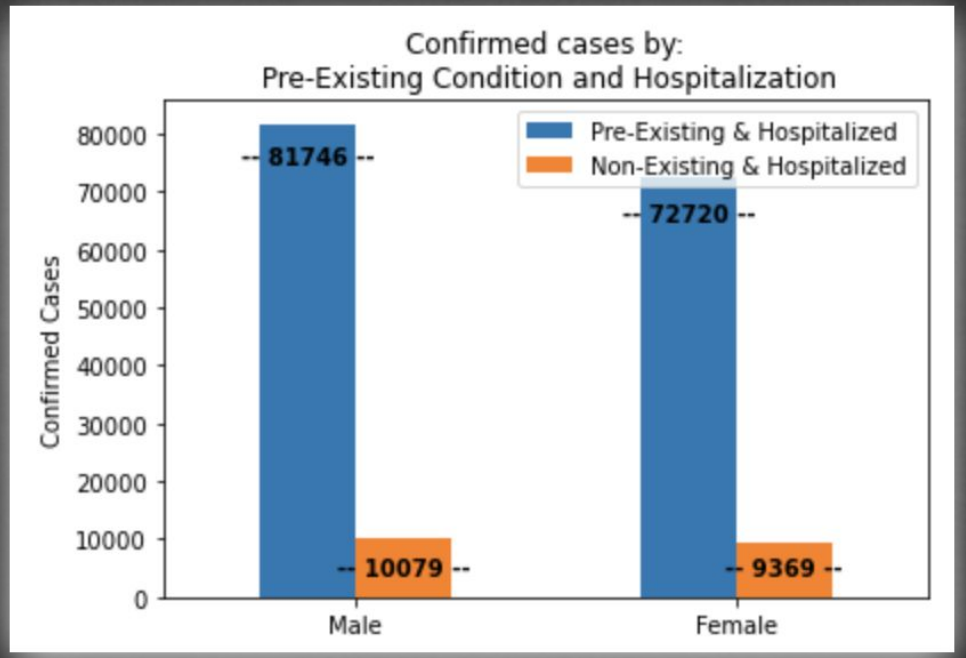


Hypothesis 3

If a person has had a pre-existing medical condition and contracts COVID-19, then they are more at risk to be hospitalized.

Discussion:

- Close similarity between male and female patients: % of those with pre-existing conditions who contracted the virus.
- In looking at our population, 56% of males who contracted the virus had pre-existing conditions.
- 57% of females who contracted the virus had pre-existing conditions.

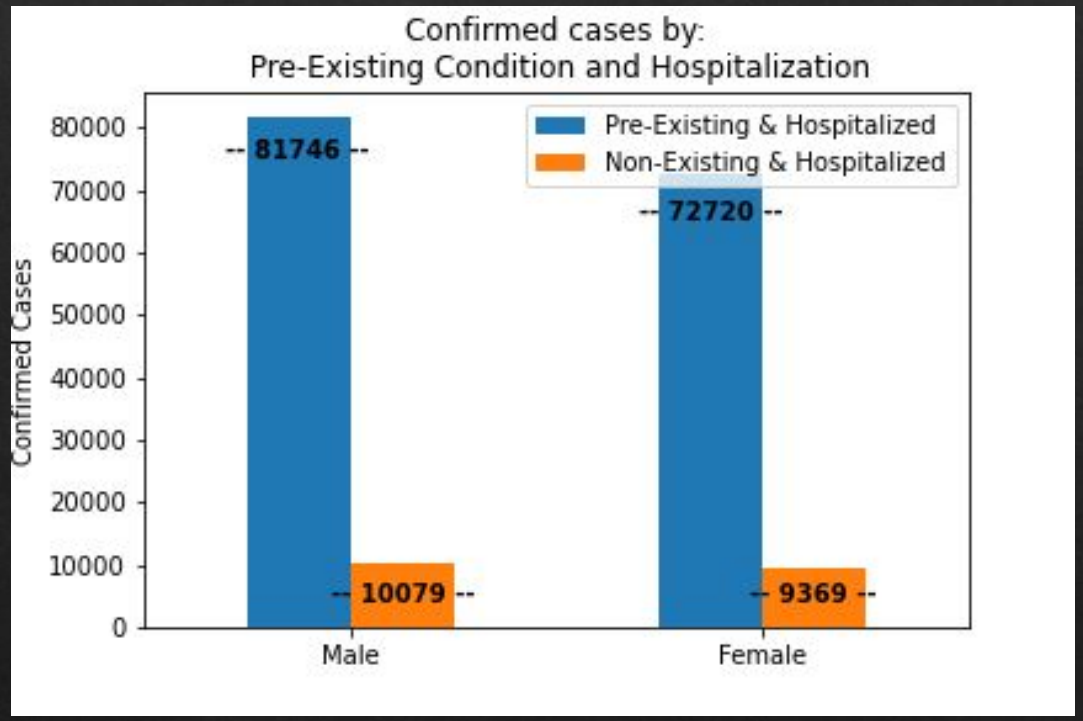



```
# Look at gender vs number of hospitalizations by pre-existing conditions
mhp = len(org_data_df.loc[(org_data_df["Gender"] == "Male") &
                          (org_data_df["Pre-Existing Condition"] == "Yes") &
                          (org_data_df["Hospitalized"] == "Yes")])
fhp = len(org_data_df.loc[(org_data_df["Gender"] == "Female") &
                          (org_data_df["Pre-Existing Condition"] == "Yes") &
                          (org_data_df["Hospitalized"] == "Yes")])

mhnp = len(org_data_df.loc[(org_data_df["Gender"] == "Male") &
                          (org_data_df["Pre-Existing Condition"] == "No") &
                          (org_data_df["Hospitalized"] == "Yes")])
fhnp = len(org_data_df.loc[(org_data_df["Gender"] == "Female") &
                          (org_data_df["Pre-Existing Condition"] == "No") &
                          (org_data_df["Hospitalized"] == "Yes")])
```

	Pre-Existing	Non-Existing
Male	259,987	204,863
Female	299,876	227,161

Hypothesis 3

If a person has had a pre-existing medical condition and contracts COVID-19, then they are more at risk to be hospitalized.



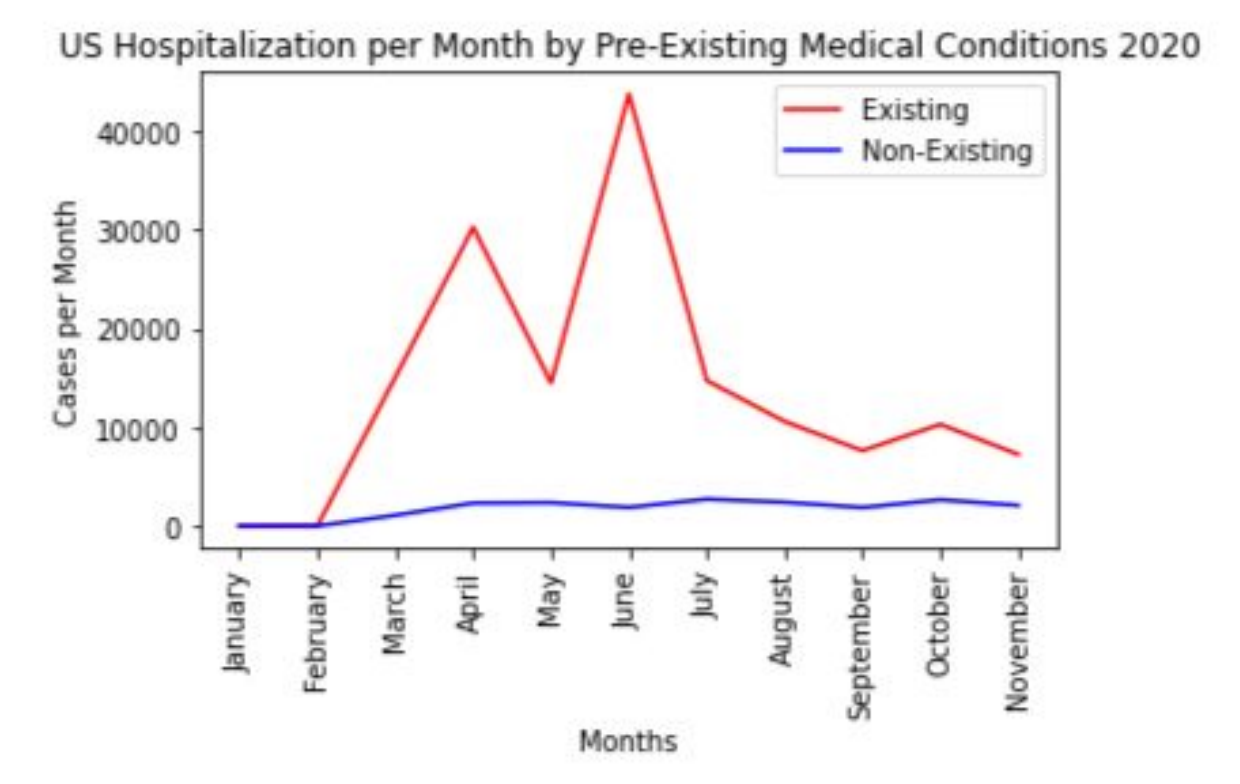
	Pre-Existing & Hospitalized	Non-Existing & Hospitalized	Hospitalized w/ Pre-Existing Condition %
Male	81,746	10,079	89.02%
Female	72,720	9,369	88.59%

Discussion:

- Looked at multiple conditions: Pre-Existing Condition & Hospitalization
- Again, we see a very close similarity between the two subsets of our population - within 1% point.

Hypothesis 3

If a person has had a pre-existing medical condition and contracts COVID-19, then they are more at risk to be hospitalized.



Discussion:

- Line chart shows our change over time; continuing to show that:
  - Individuals with a pre-existing condition are much more likely to be hospitalized
  - With this, we are not able to reject the null-hypothesis.
- Given that communities started to phase back open in late May it is not all that surprising to see such a large spike in overall hospitalizations in June.

Summary

- **Core Question:** Who is most likely to be hospitalized?
  - **Hypothesis 1:** If males are more susceptible to Covid-19, then more hospitalizations will occur among male patients.
    - **Conclusion:** Men appear to be slightly more susceptible to COVID-19, with men being hospitalized 5.6% more, but the margin is small enough to require further analysis.
  - **Hypothesis 2:** If the elderly are more susceptible to Covid-19, then more hospitalizations will occur among elderly patients.
    - **Conclusion:** The elderly appear to be much more susceptible to the effects of COVID-19, with hospitalizations sharply increasing in the 50's and above.
  - **Hypothesis 3:** If a person has has a pre-existing medical condition and contracts COVID-19, then they are more at risk to be hospitalized.
    - **Conclusion:** People with pre-existing conditions appear to be at very high risk of being hospitalized when they contract COVID-19, with 88-89% requiring hospitalization.

Post Mortem

- **Difficulties**
  - Limiting the scope, what we could realistically accomplish
  - Finding data/dataset that allowed us to look at all these demographics

- **Further Explorations**
  - What would hospitalizations look like if we compared Blue States versus Red States?
  - Which ethnicities would be more impacted by Covid-19?
  - How does Covid-19 look over time?

Questions & Answers

Resources

Archived: WHO Timeline - COVID-19. (2020, April 28). World Health Organization. https://www.who.int/news/item/27-04-2020-who-timeline---covid-19

Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Public Data Access, Summary, and Limitations (version date: December 04, 2020).