Case Study

Homeless Encounters in California Hospitals

Problem: California has 20% of the nations homeless population. Everyone deserves healthcare, so how can we continue to provide better healthcare to the homeless community in California?

Goal: Create awareness of locations with the most need. What/if any demographic traits may increase the likelihood of a hospital encounter.

Role: Data Analyst

Key Questions:

- What variables from the demographics (age, race, and sex) have the highest number of encounters?
- Are there certain counties that are hit harder with homeless encounters?
- Are the counties in which are are hit the hardest in an area where there are primary care and/or mental health practitioner shortages?
- Are the homeless more likely to have an encounter in the Emergency Department or Inpatient Hospitalization?
- How can we use this analysis to better serve homeless patients?

Tools Used:

Data

- "Hospital Encounters for Homeless Patients" Accessed from https://data.chhs.ca.gov/dataset/hospital-encounters-for-homeless-patients
- Data Dictionary: Accessed from "Hospital Encounters for Homeless Patients" Accessed from https://data.chhs.ca.gov/dataset/hospital-encounters-for-homeless-patients
- "Homelessness in USA", Accessed from https://datahub.io/gavram/homelessness#readme

Python Skills

- Importing libraries and datasets
- Descriptive analysis
- Data wrangling
- Grouping and aggregating data
- Visualisations with Python libraries
  - Choropleth Map
  - Categorical Plots
  - Regression Analysis
  - Cluster Analysis
  - Time Series Analysis
  - Correlation Matrix
- Exporting datasets

Challenges/Limitations

- I did not have a dataset listing homeless population in California over time. This would have allowed us to look for trends or seasonality to see if the homeless population has increased or decreased over time in California. I used a dataset that had the entire US homeless population and therefore did not include in the analysis as it was not specific to California. Going forward I would try web scraping to see if I could collect the necessary data.

- The hospital encounters dataset is prone to errors due to manual entry; however, data is still assumed to be reliable.

- Encounters are not representative of the population. One person may have 10+ encounters and another person may only have 1.

- Most of the data was categorical vs continuous creating a challenge when completing the cluster analysis to find additional correlations. Going forward, an alternative would be trying a cluster analysis using dummy variables on the categorical data.

Objective:

Perform exploratory and statistical analysis to derive insights on how the healthcare community can better support homeless patients as California has 20% of the nations homeless population.

Data Sourcing

Find reputable dataset which contains the following:

- > 2K observations
- > 3 categorical variables
- 3 continuous variables

Data Cleaning & Deriving New Variables

Clean and wrangle data

Gain thorough understanding of data through:

- Renaming columns
- Removing columns
- Changing data types
- Clean up inaccurate data

Filter and Aggregate Data

Obtain descriptive statistics

Data Exploration & Analysis

Create Choropleth Map showing the top 5 counties with the greatest number of homeless hospital encounters.

Exploratory Analysis using categorical plots.

Standardized data and created Time Series Regression Analysis.

Cluster Analysis

View Correlations through Heatmatrix and Correlation Matrix

Visualizations

Create visualizations in

Python using:

- Folium
- Matplotlib
- Plotly

Tableau

- Create Analytical Dashboard

Presenting Results

Utilize Tableau to present analytical dashboard through a story

Add project code to Github repository.

Tools Used:



Datasets and Code can be found on Gitbhub repository

Initial Exploration: Location

The first step after gathering and cleaning the data was to see which counties in California had the most encounters.

- Southern California and San Francisco area had the highest amount of homeless encounters.

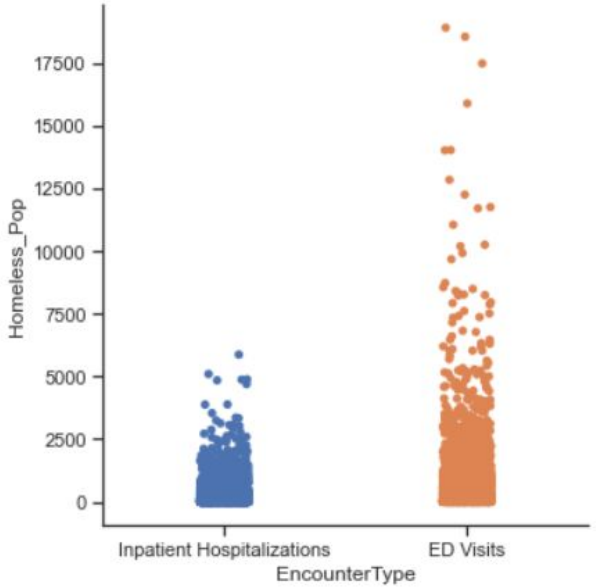
Initial Exploration: Hospital Demographics

After looking at counties, exploration moved to seeing what part of the hospital these encounters happened most.

- Emergency Department visits were more common than Inpatient Hospitalizations.

6d. Where are the homeless more likely to go for care (The Emergency Department or Inpatient Hospitalization)?

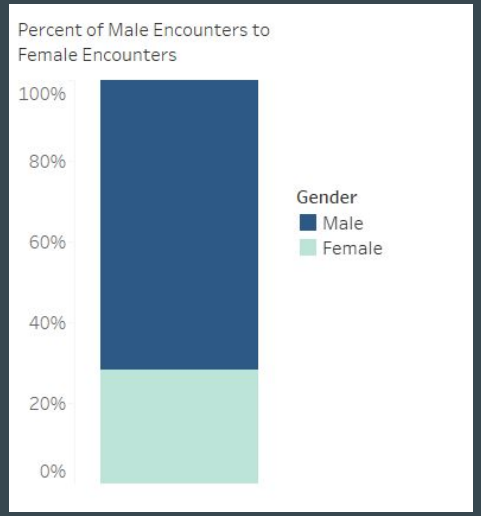
```
# location of visits
sns.set(style='ticks')
location = sns.catplot(x='EncounterType', y="Homeless_Pop", data = df2)
```

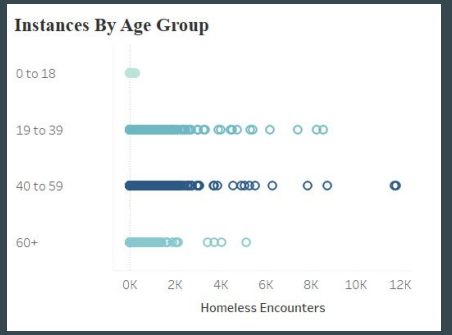


Initial Exploration: Patient Demographics

Lastly looking at the patient demographics to see if there are any trends with the patients.

- Male encounters are about 70% of visits, vs 30% for females.
- Age group 40-59 has the most encounters.
- White/Caucasian race has the most encounters out of the identified races.





Homeless Encounters by Race

White/Caucasian	401,237
Hispanic	208,008
Black	199,702
Other Race/Ethnicity	64,905
Asian/Pacific Islander	16,905
American Indian/Alaska Native	5,428

Comparing how the regression fits the training set

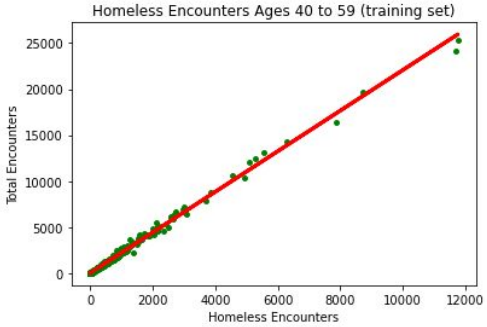
```
# Predict the training set
y_predicted_train = regression.predict(X_train)
```

```
rmse = mean_squared_error(y_train, y_predicted_train)
r2 = r2_score(y_train, y_predicted_train)
```

```
print('Slope:' , regression.coef_)
print('Mean squared error: ', rmse)
print('R2 score: ', r2)
```

```
Slope: [[2.2043555]]
Mean squared error:  34330.96339790462
R2 score:  0.9949416598816624
```

```
# visualizing the training set results
plot_test = plt
plot_test.scatter(X_train, y_train, color = 'green', s=15)
plot_test.plot(X_train, y_predicted_train, color='red', linewidth=3)
plot_test.title('Homeless Encounters Ages 40 to 59 (training set)')
plot_test.xlabel('Homeless Encounters')
plot_test.ylabel('Total Encounters')
plot_test.show()
```



- Analysis on training set: This line matches up very closely with the data! We can tell there is a strong correlation between the Homeless Encounters ages 40-59 compared to the other ages.

Digging Deeper with Linear Regression

Because the age group 40-59 had the most encounters, further research was done to see if there was a strong correlation between this age group and total encounters of other age groups.

- Yes, there is a significant difference between hospital encounters in the age group 40-59, compared to those of the other age groups.

Final Report

Recommendations:

- Priority should be given to LA county first, then Southern California and San Francisco Bay area.

- Track overarching themes of what brings patients to the hospital.
    - This allows analysts to find out if there are trends where age-appropriate resources can be provided.

- Further exploration needed on why males have more encounters than females.

- Discussion on how hospitals can be more proactive about health care needs for homeless patients.
    - This may help to reduce the amount of Emergency Department visits.

- BONUS: Ensure staff, administrators, and clinicians are physically, mentally, and emotionally trained to handle homeless patients.
    - This helps staff feel empowered and equipped to handle these challenging situations when they arise, thus providing overall better staff and patient experience.

Deliverables

- Github Repository Link
    - Project Brief
    - Datasets
    - Code/Scripts
    - Final Report
- Tableau Storyboard
- Portfolio Website