Saagnik Sarbadhikari, 50592327
Marcus Hartman, 50398874
Bharath Reddy, 50563984

Project Phase 1

1. Problem Statement (5pts, team)

Title: Forecasting the Rental Housing Economy of Various Areas Using Machine Learning.

Problem Statement: Develop a model that can accurately predict rent prices across various locations in the next few years. This will use past data along with future predictions on the economy, which can be used by various clients to find the most cost-effective location to rent.

Potential of the Project

Explain the potential of your project to contribute to your problem domain. Discuss why this contribution is crucial?

Our project aims to be useful in multiple aspects of the world. While our project isn't necessarily predicting the outcome of the economy in the next couple of years, we are aiming to get a gauge of the possibilities that the housing economy could take. This contribution aims to prove its usefulness to both current residents and people looking to move into a new area, as they will have a resource at their disposal that could aide in finding the best-fitting place to live a comfortable life economically.

2. Ask Questions (10pts, individual)

Saagnik:

(a) Could there be a larger increase in the gap of rent between a high-income state and one that is lower? Could we see people move to other states to save the money, but work remotely?

This is a significant question as it poses a greater risk ever since the rise of remote work. We have seen people already do this, i.e., moving from New York City to Buffalo for this exact reason. I am wondering if this could be seen on a much larger scale, as rent continues to rise in those higher-income states.

(b) As rent continues to rise, could we see a decline of these prices to combat multi-family apartments?

The significance of this question arises to illustrate the point of more people opting to live with multiple other people in order to save on rent in smaller buildings. Splitting rent for a three-bedroom apartment between seven people can be much more affordable than each having their own apartment—will the market react to this? What would this look like as prices keep rising? This is what an additional hypothesis can be formed to question. The best way to answer this question is to keep track of the price differences between one-bedrooms and multi-bedrooms in the future.

Marcus:

(a) Could there be a larger increase in the gap of rent between a high-income state and one that is lower? Could we see people move to other states to save the money, but work remotely?

This is a significant question as it poses a greater risk ever since the rise of remote work. We have seen people already do this, i.e. moving from New York City to Buffalo for this exact reason. I am wondering if this could be seen on a much larger scale, as rent continues to rise in those higher income states.

(b) As rent continues to rise, could we see a decline of these prices to combat multi-family apartments?

The significance of this question arises to illustrate the point of more people opting to live with multiple other people in order to save on rent in smaller buildings. Splitting rent for a three-bedroom apartment between 7 people can be much more affordable than each having their own apartment– will the market react to this? What would this look like as prices keep rising? This is what an additional hypothesis can be formed to question. The best way to answer this question is to keep track of the price differences between one-bedrooms and multi-bedrooms in the future.

Bharath:

(a) Given the higher mean and variation in rent prices in Los Angeles compared to Rochester, what implications does this have for the affordability of housing in these cities?

This question is significant as it highlights the economic disparity between high-cost living areas like Los Angeles and more affordable regions like Rochester. Understanding these implications can inform policymakers and potential renters about the sustainability of living in high-rent areas as economic conditions evolve.

(b) How does the consistency of rent prices in Rochester compared to the wider spread of rent prices in Los Angeles affect renters' choices in these cities?

This question seeks to explore the impact of rent price stability on the decision-making process of potential renters. If Rochester offers more predictable rental costs, could it become a more attractive option for individuals seeking long-term housing stability, particularly as remote work becomes more common?

3. Data Retrieval (5pts, team)

To help with individual questions that will be asked in Phase 2 and 3, our group has agreed to focus into two states per person. This will make it easier to strike our own findings, and can find different conclusions using the same model.

Saagnik:

Rural:Seattle

Urban:New York City

Marcus:

Rural: Minnesota

Urban: Texas

Bharath:

Rural:Los Angeles

Urban:Rochester

Data Retrieval

Sources used (so far):

(a) huduser.gov - 40th Percentile Rents (Historic)

(b) zillow.com - Current housing markets, estimations for an area

(c) rentcast.io - Additional market data for a zip code area

(d) insideairbnb.com - AirBnB rentals by city

Data retrieval consists of a couple calls to certain APIs. While the first two sources are free, rentcast unfortunately charges after the 31st use of the month. Therefore, we only try to do this call once and store the result in the results/ folder we have for this.

The code can be found in the associated ipynb, where the following lines are found:

For huduser.gov:

```python
# huduser.gov
huduser_key = open('tokens/huduser').readline()

huduser_MN = requests.get("https://www.huduser.gov/hudapi/public/fmr/statedata/MN",
headers={"Authorization": f"Bearer {huduser_key}"}).json()

huduser_TX = requests.get("https://www.huduser.gov/hudapi/public/fmr/statedata/TX",
headers={"Authorization": f"Bearer {huduser_key}"}).json()
```

For zillow.com:

```python
# zillow.com
zillow_rentals = pandas.read_csv("https://files.zillowstatic.com/research/
public_csvs/zori/Metro_zori_uc_sfrcondomfr_sm_month.csv?t=1728278819")
# this is home values forecasts, not rentals. but still poses significance
zillow_forecasts = pandas.read_csv("https://files.zillowstatic.com/research/
public_csvs/zhvf_growth/Metro_zhvf_growth_uc_sfrcondo_tier_0.33_0.67_sm_sa_month.csv?
t=1728278819")
```

For rentcast.com:

```
# rentcast.io
# NOTE: any request using the API key will accumulate towards the monthly limit
(31). please do not go over this

# the commented out part uses the API key you put into tokens/rentcast (req. credit
card info).
# the service is free but still requires credit card info-- if you find something
better use it. update latex and make a little note next to it and we're all good
# rentcast_key = open('tokens/rentcast').readline()

# rentcast_TX = requests.get("https://api.rentcast.io/v1/listings/rental/long-term?
state=TX", headers={"accept": "application/json", "X-Api-Key": rentcast_key})
# f = open("results/rentcast_TX", "w")
# f.write(json.dumps(rentcast_TX.json()))
# f.close()

# rentcast_MN = requests.get("https://api.rentcast.io/v1/listings/rental/long-term?
state=MN", headers={"accept": "application/json", "X-Api-Key": rentcast_key})
# f = open("results/rentcast_MN", "w")
# f.write(json.dumps(rentcast_MN.json()))
# f.close()

# loads data stored from api calls above. costs money if we go above 31 calls / month
rentcast_TX = json.loads(open('results/rentcast_TX').readline())
rentcast_MN = json.loads(open('results/rentcast_MN').readline())
```

For insideairbnb.com:

```
# airbnb
airbnb_TX_data = pandas.read_csv("https://data.insideairbnb.com/united-states/tx/dallas/
2024-08-17/visualisations/listings.csv")
airbnb_MN_data = pandas.read_csv("https://data.insideairbnb.com/united-states/mn/
twin-cities-msa/2024-06-24/visualisations/listings.csv")
```

4. Data Cleaning (10pts, team)

For each dataset, we cleaned up all of the NaNs, converted every dataset into a similar JSON, and dropped any unnecessary columns pertaining to our problem. The reason why JSON is chosen is because pandas provides methods to convert such into pandas DataFrames, which provides exceptional methods that will aid in this project. Each are separated out by their source.

For huduser.gov:

```python
# 1. cleaning hudusers.gov of unnecessary columns
def cleanCounties(countyInfo):
    columnsToDrop = ['town_name', 'metro_name', 'fips_code', 'FMR Percentile',
    'metro_name', 'statename', 'statecode', 'smallarea_status']

    for county in countyInfo:
        for n in columnsToDrop:
            if n in county:
                county.pop(n)

    return countyInfo


countyInfo_MN = cleanCounties(huduser_MN['data']['counties'])
countyInfo_TX = cleanCounties(huduser_MN['data']['counties'])
```

For zillow.com:

```python
#2. Zillow rentals: drop unnecessary columns
zillow_rentals.drop(["RegionID", "SizeRank", "RegionType"], axis=1, inplace=True,
errors='ignore')

#3. Zillow rentals: drop all that aren't for MN, TX
zillowRentals_MN = zillow_rentals.query('RegionName.str.contains("MN")')
zillowRentals_TX = zillow_rentals.query('RegionName.str.contains("TX")')

#4. Zillow forecasts: drop unnecessary columns
zillow_forecasts.drop(["RegionID", "SizeRank", "RegionType", "BaseDate"], axis=1,
inplace=True, errors='ignore')

#5. Zillow forecasts: drop all that aren't for MN, TX
zillowForecasts_MN = zillow_forecasts.query('StateName == "MN"')
zillowForecasts_TX = zillow_forecasts.query('StateName == "TX"')
```

For rentcast.com:

```python
#6. rentCast: drop unnecessary columns
def cleanRentals(rentalInfo):
    columnsToDrop = ['id', 'formattedAddress', 'addressLine1', 'addressLine2',
    'latitude', 'longitude', 'status', 'listingType', 'listedDate', 'removedDate',
    'createdDate', 'lastSeenDate', 'daysOnMarket', 'yearBuilt', 'history']

    for rental in rentalInfo:
        for c in columnsToDrop:
            if c in rental:
                rental.pop(c)

try:
    cleanRentals(rentcast_MN_json)
    cleanRentals(rentcast_TX_json)
except TypeError: # cell already ran
    pass

#7. rentCast: sort data by county, store in dictionary for easier lookup
def convertToDict(rentalInfo):
    res = dict()
    for rental in rentalInfo:
        county = rental['county']
        if county not in res:
            res[county] = []
        res[county].append(rental)
    return res
```

For insideairbnb.com:

```python
#8. airbnb: get rid of unnecessary columns
airbnb_MN_data.drop(["id", "name", "host_name", "host_id", "neighbourhood_group",
"latitude", "longitude", "last_review", "reviews_per_month", "number_of_reviews_ltm",
"license"], axis=1, inplace=True, errors='ignore')
airbnb_TX_data.drop(["id", "name", "host_name", "host_id", "neighbourhood_group",
"latitude", "longitude", "last_review", "reviews_per_month", "number_of_reviews_ltm",
"license"], axis=1, inplace=True, errors='ignore')

#9. airbnb: get rid of duplicates
airbnb_MN_data.drop_duplicates()
airbnb_TX_data.drop_duplicates()

#9. airbnb: get rid of rows where price is not available
airbnb_MN_data.dropna(subset=['price'], how='all', inplace=True)
airbnb_TX_data.dropna(subset=['price'], how='all', inplace=True)

#10. airbnb: sort by room_type, convert to dictionary
def convertToDict_BNB(df : pandas.DataFrame):
    res = dict()
    for _, row in df.iterrows():
        roomType = row['room_type']
        if roomType not in res:
            res[roomType] = []
        res[roomType].append(row)
    return res

airbnb_MN = convertToDict_BNB(airbnb_MN_data)
airbnb_TX = convertToDict_BNB(airbnb_TX_data)
```

5. Exploratory Data Analysis (20pts, individual)

Saagnik:

(a) Out of the two states I have chosen, if I am looking at the data of New York compared to Seattle, I will see a higher mean and more variation of rent in New York because it is more economically active than Seattle.

The following code is used to produce this example (data cleaning did most of the work here):

```
# EDA Question 1: Rent Comparison between NYC and Seattle

# Extracting the most recent rent data from Zillow rentals
nyc_rent = zillowRentals_NY['2024-08-31']
seattle_rent = zillowRentals_WA['2024-08-31']

# Create a DataFrame for rent comparison
rent_data = pd.DataFrame({
    "New York City": nyc_rent,
    "Seattle": seattle_rent
})

# Plotting boxplot for rent comparison
plt.figure(figsize=(10, 6))
rent_data.plot(kind='box', title="Rent Comparison: New York City vs Seattle (August 2024)
", ylabel="Rent (USD)")
plt.show()
```

Which produces this graphic:
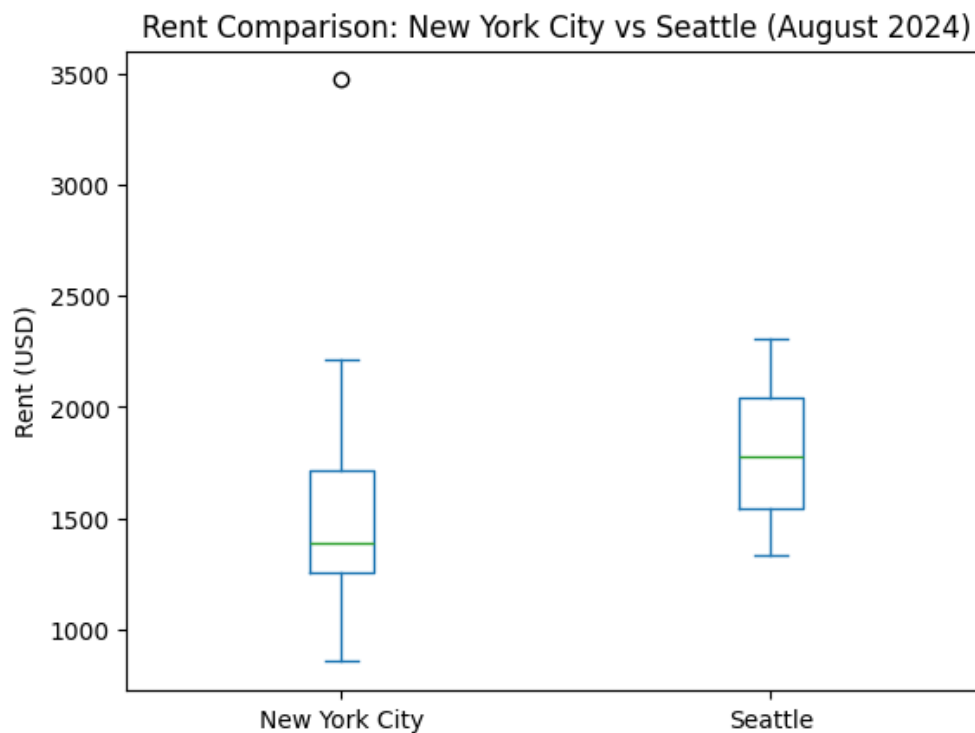
Rent Comparison: New York City vs Seattle (August 2024)

I tried to put an explanation in the code for this hypothesis. However, this will be expanded upon below.

The results show that New York experiences a higher mean rent and greater variation compared to Seattle, indicating a more dynamic rental market. This supports the hypothesis that economic activity influences rental prices and their distribution. Understanding these differences is crucial for predicting future rental trends and housing policy impacts.

(b) If New York's major boroughs have high rent, then there will be much lower rent outside of these boroughs because there is a lower standard of living comparatively.

Below is the code that will create the graphic:

```python
# EDA Question 2: Airbnb Listings Price vs. Room Type in New York City

# Grouping by room_type and calculating the mean price
airbnb_nyc_prices = airbnb_NYC_data.groupby('room_type')['price'].mean().reset_index()

# Plotting bar chart for average price by room type
plt.figure(figsize=(10, 6))
plt.bar(airbnb_nyc_prices['room_type'], airbnb_nyc_prices['price'], color='skyblue')
plt.title("Average Airbnb Prices by Room Type in NYC (2024)")
plt.xlabel("Room Type")
plt.ylabel("Average Price (USD)")
plt.xticks(rotation=45)
plt.show()
```

The graphic:
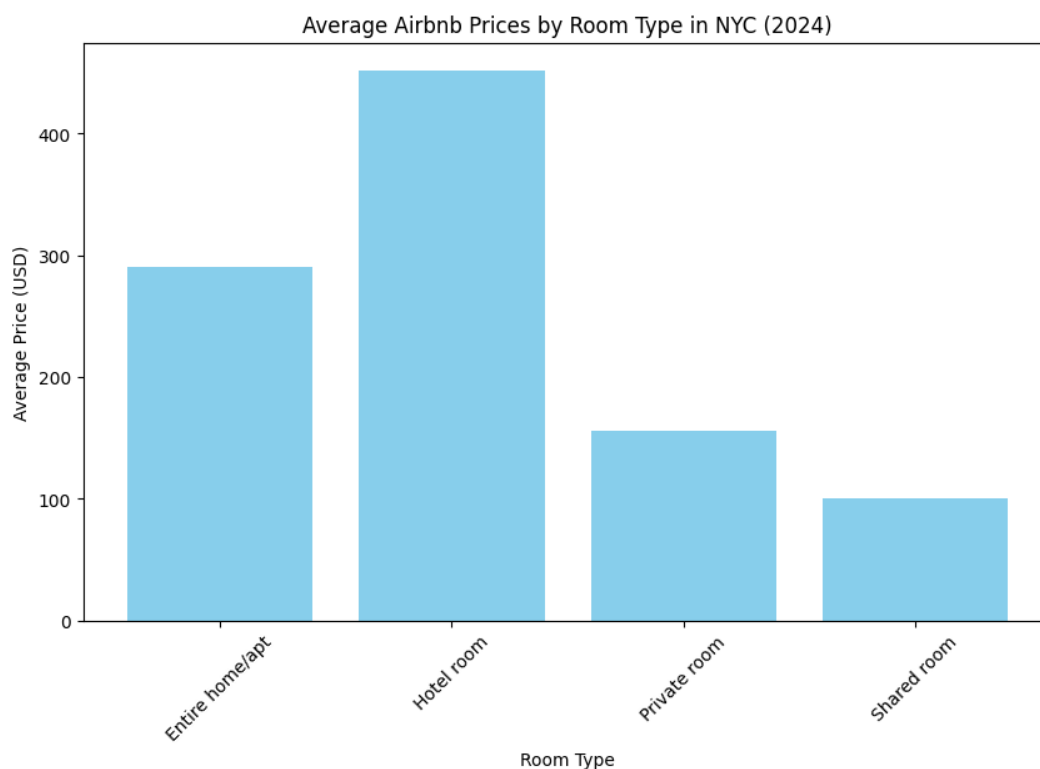
At first glance, this graphic may not seem to carry significant weight, but it reveals important

insights about rent distribution and government assistance. The lines of efficiency indicate hard boundaries that illustrate the necessary rent efficiency compared to population. The upper limit approaches $3,000 per month, while the lower end hovers around 1,200$.

Government assistance plays a pivotal role in maintaining rental efficiency, ensuring that landlords do not perceive their properties as financial losses. This assistance often translates to an additional $500 to 1,000$, depending on various economic factors.

The upper boundary signifies what tenants consider 'reasonable' rent within their economic means. Rental efficiency is a crucial metric that reflects this economic reality, and more detailed formulas can be found on the HUD user website.

In the context of predicting trends in future phases, this hypothesis reinforces the idea that economic conditions and governmental interventions significantly impact the rental landscape. Therefore, to achieve accurate predictive outcomes, it is vital to simulate an increasing gap in rental efficiency across different counties, accounting for economic growth disparities.

Marcus:

(a) Out of the two states I have chosen, if I am looking at the data of Texas compared to Minnesota, I will see a higher mean and more variation of rent because Texas is more economically active than the other.

The following code is used to produce this example (data cleaning did most of the work here):

```python
# zillowrentals_MN contains historic data of all cities in the associated state
MN_current = zillowRentals_MN['2024-08-31']
TX_current = zillowRentals_TX['2024-08-31']

# combine then plot. boxplot will show all the descriptors we need
pandas.DataFrame({"Minnesota" : MN_current, "Texas" : TX_current}).plot(kind='box',
title="Minnesota Rent against Texas Rent, 2024")

# in terms of the hypothesis, this is proven true.
# the average of TX rent is much higher than the MN rent, as expected. The variation of
rent can be shown by looking at the differences between IQ1 and IQ3, which are much
farther apart in TX than MN.
```

Which produces this graphic:
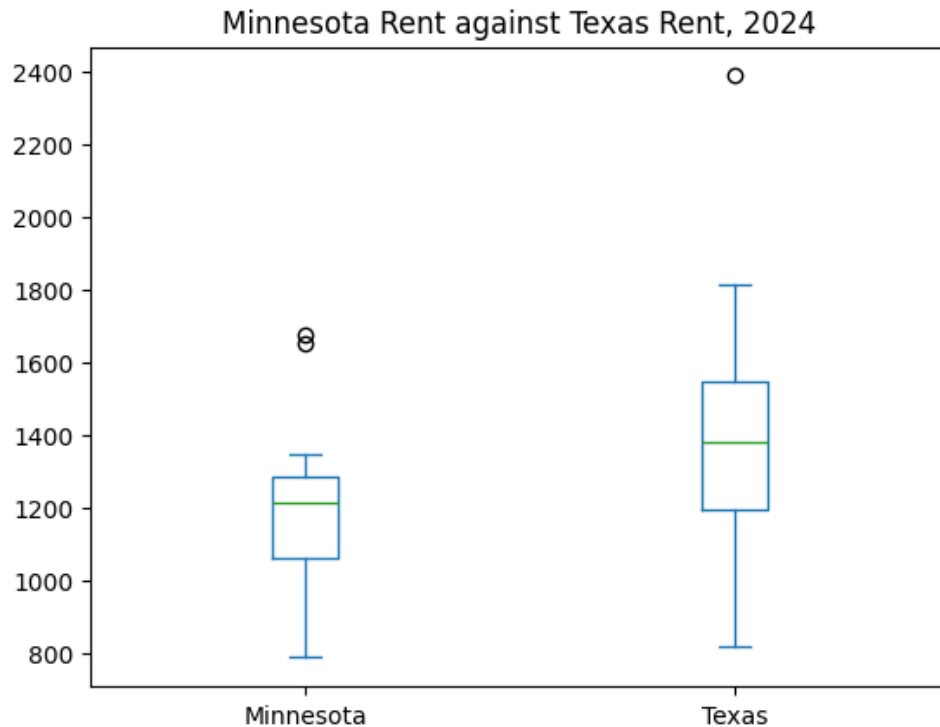
Minnesota Rent against Texas Rent, 2024

I tried to put an explanation in the code for this hypothesis. However, this will be expanded upon below.

Minnesota experiences both a lower deviation and mean, signifying a much cheaper rent economy than Texas. This proves the hypothesis correct, and yields results that will be used in the later phases. The use of this stems from understanding the different economies in place, and how additional steps will need to be taken in order to predict economic growth in different areas.

(b) If Minnesota's major cities have high rent, then there will be much lower rent outside of these cities because there is a lower standard of living comparatively.

Below is the code that will create the graphic:

```
# 2. If Minnesota's major cities have high rent, then there will be much lower rent
outside of these cities because there is a lower standard of living comparatively.

# the best source for this will be huduser for county rent efficiency
# get demographics by county (another data source, specific to the problem)
demographics_MN = pandas.read_csv('data/Demographics_MN.csv')
pop = demographics_MN.get(['pop2024', 'name'])
pop.rename(columns={'name': 'county_name', 'pop2024': 'Population'}, inplace=True)

# get counties into dataframe
counties_MN = pandas.DataFrame(pandas.json_normalize(countyInfo_MN))
counties_MN = counties_MN.merge(pop, on='county_name') # merge the demographics and
county info

# plot in scatterplot: (Efficiency, Population)
print(counties_MN)
counties_MN.get(['county_name', "Efficiency", "Population"]).plot(kind='scatter',
x='Efficiency', y="Population", logy=True, title="Counties' Efficiency of Rent versus
Population, 2024")
```

The graphic:
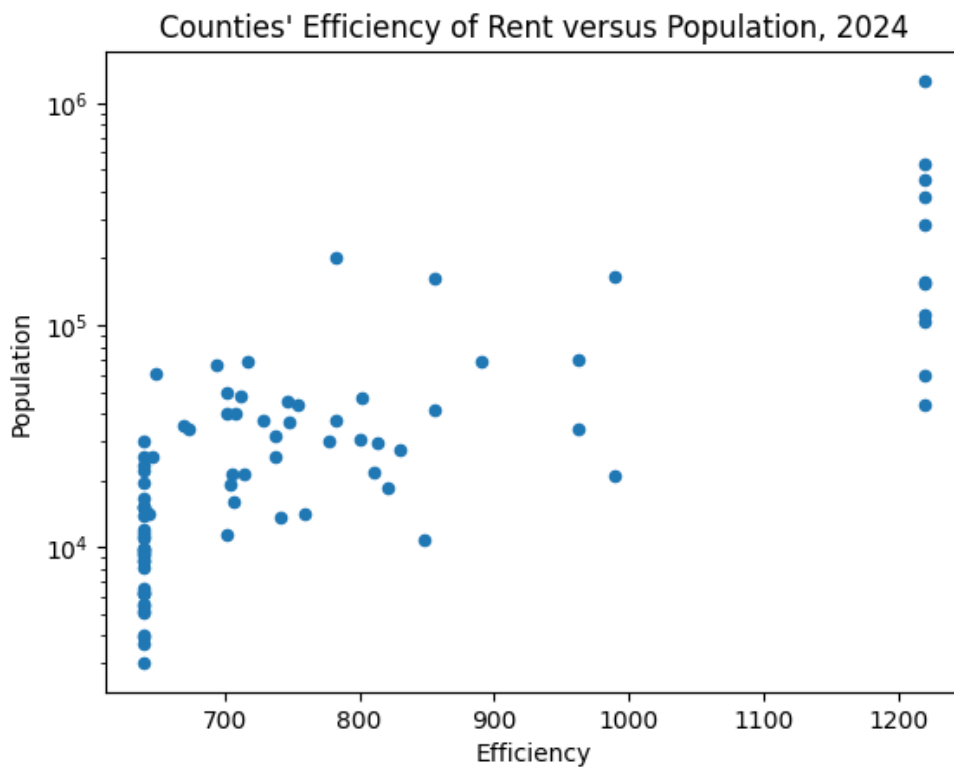
Although at first this graphics doesn't seem to have any significance, a little bit of forward thinking into government assistance is necessary. The lines of efficiency represent hard boundaries that demonstrate the limits of necessary rent efficiency compared to population. On the higher end of this, we see this cap off at slightly above 1200 per month. On the lower end, this is closer to 600.

Government assistance is used to subsidize the rent efficiency in order for the owners to not consider the property a loss. This is what the line is, and often is around 300-550 additional provided (based on income and other factors).

On the higher end, the barrier symbolizes what people will deem as 'reasonable' rent given the economic state that they can properly afford. Rental efficiency does take this into account, and the formula can be found on the huduser.gov website.

In terms of phase 2 and phase 3, this has a similar outcome compared to my first hypothesis. This shows the significance the economy and government assistance has on the rental climate. We cannot expect the lower efficient counties to not have increase government assistance while not experiencing some form of exponential economic growth in the higher efficient counties. In terms of predicting something of this nature, an increased gap will have to be simulated in order to garner accurate results of predictions.

Bharath:

(a) Given the higher mean and variation in rent prices in Los Angeles compared to Rochester, what implications does this have for the affordability of housing in these cities?

This question is significant as it highlights the economic disparity between high-cost living areas like Los Angeles and more affordable regions like Rochester. Understanding these implications can inform policymakers and potential renters about the sustainability of living in high-rent areas as economic conditions evolve.

(b) How does the consistency of rent prices in Rochester compared to the wider spread of rent prices in Los Angeles affect renters' choices in these cities?

This question seeks to explore the impact of rent price stability on the decision-making process of potential renters. If Rochester offers more predictable rental costs, could it become a more attractive option for individuals seeking long-term housing stability, particularly as remote work becomes more common?