Modern Data Management (46-884)
Mini-2, 2024
Lab 1 Analysis Questions

For this part of the lab, you need to retrieve and analyze data in the assigned datasets to generate a compelling answer to the question posed. Your answer needs to be presented in a clear and compelling way that makes your conclusions and insights clear to the audience identified for each question. In your answers for all of these questions, assume that the person reading your report on your analytic findings does not have a strong background in statistics or databases and present your findings and insights accordingly.

For each of the three questions, retrieve and analyze data from the specified dataset to answer the question posed. We will use the rubric provided for the lab on Canvas to evaluate how effectively you have analyzed the data, answered the question, and presented your findings.

Your answer should follow the template provided for each question, explicitly addressing the questions posed on that template. **Use no more than one page to present your findings for each question**. In addition to the one page presentation of your findings and insights, you need to include the code you used to retrieve and analyze the data in an appendix at the end of the submitted document. That appendix does not count toward your 1 page analytic answer. Do not include pages and pages and pages of data returned by your queries in your appendix. Your analytic results should synthesize what that data tells you about the answer to the question posed, and the graders should be able to retrieve the data used with the code provided if necessary to grade or test your answers.

Deliverable format:
Submit your answers to all three of your questions to Canvas in a single PDF document. Use a fresh page to answer each question, and complete your answer to that question before the end of that page. If you can provide a clear, compelling, and well supported answer to the question in less than a full page, then do so. Don’t use a full page if you can provide an answer that meets all of the criteria in less than that. Attach any appendices to the end of the document, after the three pages answering the questions.

Question1: Airbnb dataset

Problem Definition: The most affordable rental market is measured using a composite Affordability Score, which combines six key metrics with assigned weightages with lower scores indicate more affordable markets:

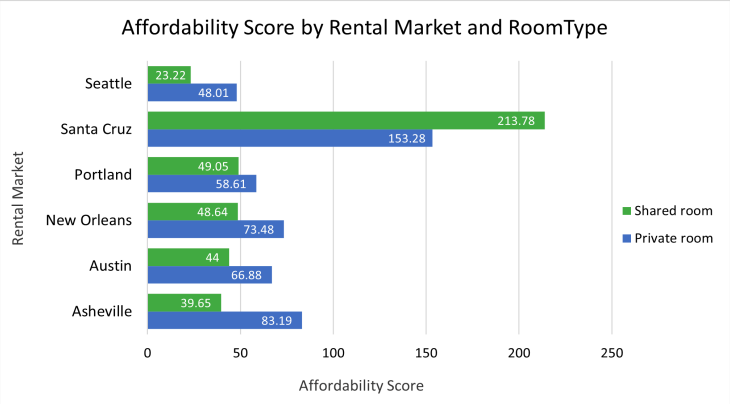
$$\textbf{\textit{Affordability Score}} = 0.3 \times \textit{Average Adjusted Price} + 0.1 \times \textit{Availability Rate} + 0.1 \times \textit{Price Per Person} + 0.2 \times \textit{Cost Efficiency} + 0.2 \times \textit{Price Per Bedroom} + 0.1 \times \textit{Minimum Nights}$$

Metrics and Calculations: *Average Adjusted Price* is the Average nightly rate after discounts, reflecting the main affordability factor. *Availability Rate* shows percentage of available days for listings. *Price Per Person* is the price adjusted by the number of people a listing accommodates. *Cost Efficiency* displays review score relative to price per person, indicating value for money. *Price Per Bedroom* is the price divided by the number of bedrooms. Useful for per-bedroom affordability. *Minimum Nights* entails the minimum stay requirement, which can impact booking flexibility & overall cost.

Top 4 Market Rentals with increasing Affordability Score:

Rental Market	Room Type	Avg Adjusted Price	Availability Rate	Price Per Person	Cost Efficiency	Price Per Bedroom	Avg Min Nights	Affordability Score
Seattle	Shared room	**29.88**	43.17%	**29**	**3.14**	**29.3**	9.196	**23.22**
Asheville	Shared room	61.05	41.23%	62	1.56	62.5	1.375	39.65
Austin	Shared room	68.57	41.78%	67	1.37	67.96	13.729	44
Seattle	Private room	76.71	51.77%	73	1.29	70.28	17.273	48.01

Limitations: We assume shared & private rooms are cheaper, but this may not be true for all markets. Moreover, affordability is influenced by the local cost of living, which complicates comparisons across cities. We also assume that customers are comfortable with the average minimum stay of Seattle, but preferences for shorter stays could make other cities, like Asheville, more attractive. Furthermore, the analysis excludes the broader rental market, which may impact Airbnb pricing and affordability.



Conclusion: Seattle stands out as the most affordable market with an **Affordability Score of 23.22 for shared rooms and 48.01 for private rooms.** This indicates Seattle as the best choice for budget-conscious travelers seeking affordable Airbnb accommodations.

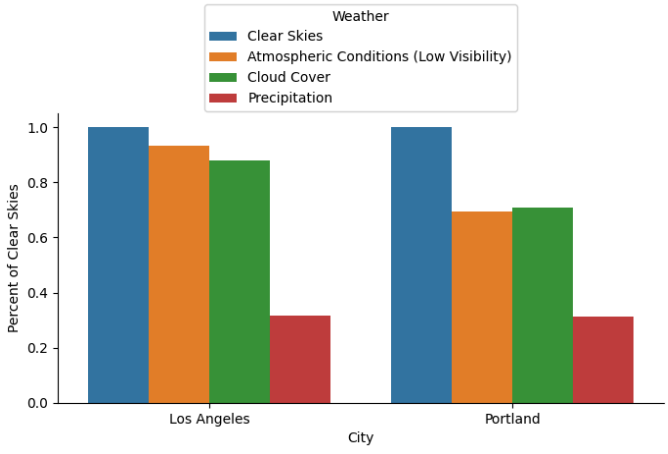
Question 2: Bikeshare dataset - Aasdeep
Analysis of Weather Impact on Bike Rentals in Los Angeles and Portland

Problem Definition: Between Los Angeles and Portland, we wanted to investigate which metro area sees the lowest proportional activity during weather conditions that are not "clear skies" to determine which bike rental business is "most affected" by weather.

Defining Data & Metrics:
To assess the impact of weather on bike rentals, we used the following quantitative metrics:

1. **Weather Conditions:** To get a representative reading of each day, we took the weather reading from noon of each day in Portland and Los Angeles.
2. **Rental Comparison:** In order to compare datasets, we only included rentals from 2017 as to limit the dataset and ensure that the data came from the same year to remove any bias in the data from different years.
3. **Total Rentals by Weather Condition**: Calculated the total rentals under different weather categories to determine the level of rental activity for each condition. The four categories we determined were from the Weather Description *column*:
    a. *'Clear Skies' - Any value equalling 'Clear Skies'*
    b. *'Cloud Cover' - Any value equalling 'Clouds'*
    c. *'Precipitation' - Any value equalling 'Snow', 'Rain' or 'Drizzle'*
    d. *'Low Visibility' - Any value equalling 'Smoke', 'Fog', 'Mist', 'Haze', or 'Dust'*
4. **Average Number of Rentals by Weather Condition**: For each metro area, we calculated the average daily rentals for each weather category, so we could compare if the number of rentals on days with precipitation decreased compared to clear days.
5. **Proportion of Clear Sky Rentals:** For each weather category, we divided its average daily rentals by the average for the clear skies category. For example:
    Average Number of Rentals (Precipitation) / Average Number of Rentals (Clear Skies)

Proportion of Clear Skies Daily Rentals



Key Findings and Comparison:
From the data, it's evident that **Portland's bike rental activity is more affected by weather than Los Angeles**. More specifically, Portland's bike rental activity is more affected by cloud coverage and low visibility compared to Los Angeles. However, when there is precipitation, both rental markets see a similar significant decrease in bike rentals.

Question 3: US Demographics dataset - Visvaant

Standards and Measurements for Classifying Population Density
To categorize counties as densely, moderately, or sparsely populated, the population density of each county in 2011 is determined using the subsequent formula:

Population Density = Total Population / Area in Sq. Miles

After analyzing the density distribution throughout Pennsylvania's counties, counties were categorized into three groups based on percentiles:

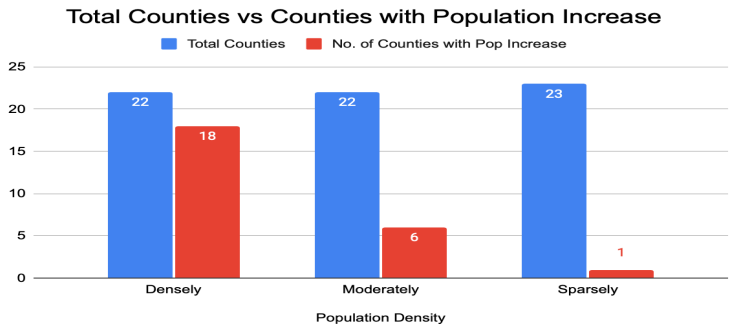
1) **Sparsely inhabited:** Counties that have a population density percentile ranking below 33.33% (lowest one-third). This is approximately equal to 83 individuals per square mile or greater.
2) **Moderately populated:** Counties that fall within a population density percentile rank of 33.33% to 66.67% (the central third). This equates to roughly between 83 and 227 individuals per square mile or higher.
3) **Densely populated:** Counties that fall in the top one-third with a population density percentile rank exceeding 66.67%. This relates to roughly. 227 individuals per square mile or greater.

To calculate the proportional population change for counties across the different population densities, we took the average of each county's percent change in population for each county or:

Average [% Change in Population] by Population Density

The following table represents the population trends in counties that belong the the above category

Population Density	Avg Percent Change in Pop
Densely	3.249
Moderately	-1.781
Sparsely	-4.85



Summary:
From 2011 to 2021, only the densely populated counties in Pennsylvania saw an increase in population, with around 86% of them exhibiting population growth and an average increase rate of 3.69%. In comparison, the counties with moderate and low populations experienced decreases, with merely 17.39% and 9.09% of counties in these categories showing any growth, respectively. The typical population shift for counties with moderate populations was -2.91%, whereas counties with low populations experienced an average decline of -4.26%. This suggests that population increase was focused in highly populated regions, whereas sparsely populated areas mostly experienced reductions in population.

Appendix:
Provide the code you used to do your analysis and calculate your metrics in this appendix.

Question 1 code:

```
/*MDM Lab 1, Part 2: Airbnb Data*/
SELECT
   MetroArea.CoreCity AS RentalMarket,
   Listing.RoomType AS RoomType,
   -- Average prices from Calendar Table
   ROUND(AVG(Calendar.AdjustedPrice), 2) AS AvgAdjustedPrice,
   ROUND(AVG(Calendar.ListedPrice), 2) AS AvgListedPrice,
   -- Availability Rate (percentage of days available)
    ROUND(
      (count( case when Calendar.IsAvailable = 1 then 1 end) + 0.0)
      / COUNT(Calendar.Date) * 100, 2) || '%' AS AvailabilityRate,
   -- Average Price Per Person
   ROUND(sum(Listing.Price)
      / count(Listing.Accommodates), 2) as PricePerPerson,
   -- Review Score Rating
   ROUND(AVG(Listing.ReviewScoresRating)/ROUND(sum(Listing.Price)
      / count(Listing.Accommodates), 2) , 2) AS CostEfficiency,
   -- Additional metrics from Listing Table
   ROUND(AVG(Listing.Price / NULLIF(COALESCE(Listing.Bedrooms, 1), 0)), 2) AS
PricePerBedroom,
   AVG(Listing.MinNights) AS AvgMinNights,
   -- Calculating the affordability score
   (ROUND(
      0.3 * AVG(Calendar.AdjustedPrice) +
      0.1 * ROUND((COUNT(CASE WHEN Calendar.IsAvailable = 1 THEN 1 END) * 1.0 /
COUNT(Calendar.Date)) * 100, 2) +
      0.1 * (SUM(Listing.Price) / NULLIF(SUM(Listing.Accommodates), 0)) +
      0.2 * ROUND(AVG(Listing.ReviewScoresRating) / NULLIF((SUM(Listing.Price) /
NULLIF(SUM(Listing.Accommodates), 0)), 2), 2) +
      0.2 * ROUND(AVG(Listing.Price / NULLIF(COALESCE(Listing.Bedrooms, 1), 0)), 2) +
      0.1 * AVG(Listing.MinNights)
   , 2)) AS AffordabilityScore
FROM
   Listing
   JOIN Calendar ON Listing.ListingID = Calendar.ListingID
   JOIN MetroArea ON Listing.MetroID = MetroArea.MetroID
   WHERE
      Listing.RoomType IN ('Private room', 'Shared room')  -- Filter for affordable room types
      AND Listing.Price IS NOT NULL
-- Grouping by City and RoomType for a market-level view of affordability
GROUP BY
   MetroArea.CoreCity, Listing.RoomType
ORDER BY
   AffordabilityScore ASC;  -- Ordering by affordability
```

Question 2 code:

```
-- Drop the temporary tables if they already exist
DROP TEMPORARY TABLE IF EXISTS T_Daily_Rentals;
DROP TEMPORARY TABLE IF EXISTS T_Daily_Weather;

-- Step 1: Create a temporary table for daily rentals, aggregated by date and metro area
CREATE TEMPORARY TABLE T_Daily_Rentals AS (
    SELECT
        r.StartDate,
        r.MetroID,
        COUNT(r.TripID) AS DailyRentalCount
    FROM
        Rentals r
    WHERE
        r.MetroID IN (2, 4)  -- Metro areas of interest
        AND r.StartDate BETWEEN '2017-01-01' AND '2017-12-31'
    GROUP BY
        r.StartDate, r.MetroID
);

-- Step 2: Create a temporary table for daily weather data with new categories
CREATE TEMPORARY TABLE T_Daily_Weather AS (
    SELECT
        wd.DateOfReading AS WeatherDate,
        wd.MetroID,
        CASE
            WHEN wd.WeatherDescriptionMain = 'Clear' THEN 'Clear Skies'
            WHEN wd.WeatherDescriptionMain = 'Clouds' THEN 'Cloud Cover'
            WHEN wd.WeatherDescriptionMain IN ('Snow', 'Rain', 'Drizzle') THEN 'Precipitation'
            WHEN wd.WeatherDescriptionMain IN ('Smoke', 'Fog', 'Mist', 'Haze', 'Dust') THEN 'Atmospheric
Conditions (Low Visibility)'
            ELSE 'Other'
        END AS WeatherCategory
    FROM
        Weather wd
    WHERE
        wd.MetroID IN (2, 4)  -- Metro areas of interest
        AND wd.HourOfReading = 12  -- Using noon as representative weather data for the day
);

-- Final query: Aggregated rentals by weather category and metro area
SELECT
    ma.CoreCity AS MetroArea,
    w.WeatherCategory,
    AVG(r.DailyRentalCount) as AvgRentals
FROM
    T_Daily_Rentals r
JOIN
    T_Daily_Weather w ON r.MetroID = w.MetroID AND r.StartDate = w.WeatherDate
JOIN
```

```
    MetroArea ma ON r.MetroID = ma.MetroID
WHERE
    ma.CoreCity IN ('Los Angeles', 'Portland')  -- Ensuring correct metro areas
GROUP BY
    ma.CoreCity, w.WeatherCategory
ORDER BY
    ma.CoreCity, w.WeatherCategory;
```

Question 3 code:

```
CREATE TEMPORARY TABLE pop_den AS
    SELECT
        l.county,
        SUM(a.pop2011) AS pop2011,
        SUM(a.pop2021) AS pop2021,
        (SUM(a.pop2011) / SUM(l.area_land_sq_miles)) AS density_2011,
        (SUM(a.pop2021) / SUM(l.area_land_sq_miles)) AS density_2021
    FROM
        acs_estimates a
    LEFT JOIN
        location l ON a.zip = l.zip
    WHERE
        l.state = 'PA'
    GROUP BY
        l.county
;

SELECT
    density,
    COUNT(*) AS "Total Counties",
    COUNT(CASE WHEN pop_change_percent > 0 THEN 1 END) AS "No. of Counties with Pop
Increase",
    ROUND((COUNT(CASE WHEN pop_change_percent > 0 THEN 1 END) * 100.0) / COUNT(*), 2)
AS "Percentage of Counties with Pop Increase",
    ROUND(AVG(pop_change_percent), 2) AS "Average Pop Change"
FROM (
    SELECT
        *,
        (pop2021 - pop2011) AS popDiff,
        ROUND(((pop2021 - pop2011) * 1.0 / pop2011) * 100, 2) AS pop_change_percent,
        PERCENT_RANK() OVER (ORDER BY density_2011) * 100 AS popden2011_percentile,
        CASE
            WHEN (PERCENT_RANK() OVER (ORDER BY density_2011) * 100) <= (100.0 / 3) THEN
'Sparsely'
            WHEN (PERCENT_RANK() OVER (ORDER BY density_2011) * 100) > (200.0 / 3) THEN
'Densely'
            ELSE 'Moderately'
        END AS density
    FROM
```

```
        pop_den
) source
GROUP BY
    density;
```