STATISTICAL ANALYSIS OF POPULATION & EMPLOYMENT ACCESS IN CENTRAL EGYPT

Team NumeriTechne

Belal Darwish	**91240216**
Ibrahim Sherif	91240055
Ahmed Amr	91240104
Ahmed Gamal	91240071
Mohamed Hosny	9230757
Amr Salah	91240543
Mahmoud Khaled	91240710

Table of Contents:

Table of Figures:

Problem Statement:

Greater Cairo area is the most densely populated area in Egypt, and this creates a major problem which is the “Spatial Mismatch”. The Spatial Mismatch is where residents in densely populated peripheral areas may lack efficient access to employment centers. This study will investigate the gap between where people live in the Greater Cairo Region and their ability to reach job opportunities using public transportation.

Background:

Study Design:

This project is an observational analysis. We will use the provided data to evaluate different inferential statistical relationships and accompany the findings with interpretable figures and mathematical models.

Response Variable:

The primary response variables are:

- **jobs_total:** Total number of employment opportunities located within each subsector.
- **jobs_total_60:** Total number of employment opportunities accessible within 60 minutes of commuting by public transportation, per subsector.
- **jobs_perc_60:** Share of employment opportunities accessible within 60 minutes by public transportation, expressed as a percentage of all employment opportunities available to each subsector.

Factors:

Controlled**:**

- Geographic Scope: The study is limited to the administrative boundaries of the Greater Cairo Region as defined in the dataset.

Uncontrolled**:**

- Population Size (pop_2018_c): The total number of residents in a sub-sector.
- District Area (area_km): The physical size of the sub-sector in square kilometers.
- Local Job Concentration (jobs_total): The number of jobs located within the sub-sector itself.

Data Characteristics:

The geographic scope covered in the study is restricted to the Greater Cairo Region. The primary challenges to account for are the relatively small sample size (with respect to the large population density of the Greater Cairo Region) and the presence of missing values and the imbalance of the (num) response variable.

Research Questions and Statistical Questions:

1. Is there a statistically significant difference in the mean percentage of accessible jobs (jobs_perc_60) between districts classified as "Central" versus those classified as "Inner" or "Outer"?
2. To what extent can the accessibility of a district (classified as High vs. Low Access) be accurately predicted based solely on its population size (pop_2018_c) and local job count (jobs_total)?
3. Is there a statistically significant positive linear correlation between a district's population size (pop_2018_c) and its total number of jobs (jobs_total) across the Greater Cairo Region?
4. Does the governorate of a district (name_gov) significantly influence the variability of employment opportunities?

Data needed for the analysis:

Dataset Used: Population & Employment Access by District in Cairo.

Variables Description: The dataset consists of 654 instances, each described by the 15 variables listed in the next page.

Variable Name	Type	Description	Units
ogc_fid	Integer	Open Geospatial Consortium Feature ID (a unique internal number generated by the database to track this specific element).	
gid	Integer	Global Identifier (Another unique ID used to identify this specific geographic feature in the dataset).	
name_gov	Categorical	Name of the government	
zoning_tfc	Categorical	Zoning Classification within each government.	
name_tfc	Categorical	Name of the district	
name_citya	Categorical	Name of the city	
sec_name	Categorical	Name of the sector	
ssec_name_	Categorical	Name of subsector	

ssec_nam_1	Categorical	Name of subsector in English	
name_compound	Categorical	Compound name (like neighborhood name)	
area_km	Integer (decimal)	Area of each subsector.	Km^2
jobs_total	Integer	Total number of employment opportunities located within each subsector.	
jobs_perc_60	Integer (decimal)	Share of employment opportunities accessible within 60 minutes by public transportation, expressed as a percentage of all employment opportunities available to each subsector.	
pop_2018_c	Integer	Population of each subsector as of 2018.	
jobs_total_60	Integer (decimal)	Total number of employment opportunities accessible within 60 minutes of commuting by public transportation, per subsector.	

Table 1. Variables Table

Descriptive Statistics:

Bar plot:

- From the bar plot shown in figure 1, we can see that the central zones (zoning_tfc) occupy more than a staggering 40% of jobs opportunities that are accessible within 1 hour by public transportation (jobs_perc_60), leading to the belief that people who live in Central zones have much more opportunities than people who live in Inner or Outer zones.

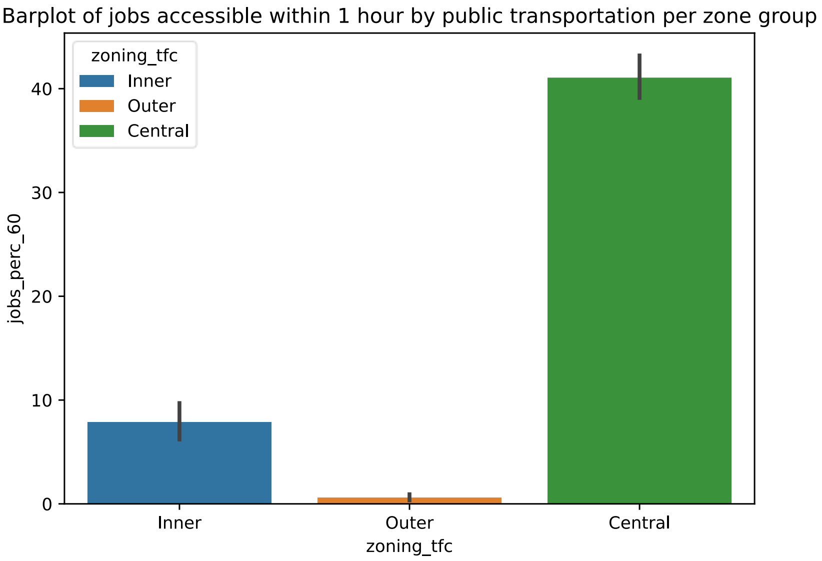


Figure 1. Bar plot of (jobs_perc_60) & (zoning_tfc)

Box plot:

- The box plot in figure 2 has been logarithmically scaled from the y-axis due to the huge number of job opportunities to represent, leading to the bottom whiskers of each plot (Q1-1.5IQR) being cut off, due to the (jobs_total) having a 0 value in each governate.
- The according figure shows that:
- Cairo has the biggest median of jobs, suggesting it is the dominant governate.
- Giza having the biggest quartile range, suggesting it has the most diverse number of jobs in its vicinity, with a good number of outliers.
- Sharkia's clear disadvantage to accessibility of jobs due to its lower number of total jobs in its grounds.

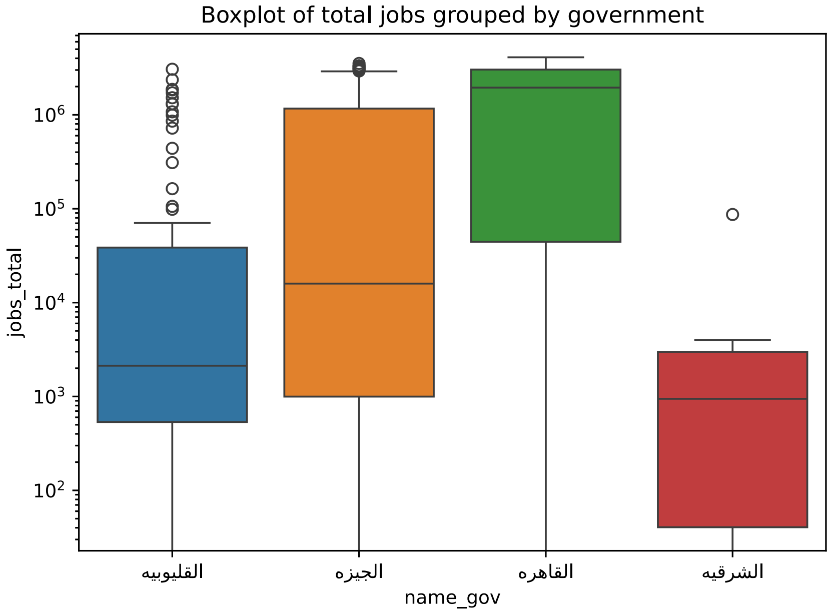


Figure 2. Box plot of (jobs_total) & (name_gov)

Histograms:

- Figure 3 shows two histograms of jobs accessible within 1 hour by public transportation (jobs_perc_60) as the histogram frequency is represented by percentages.
- Figures 3.1 and 3.2 are the same histogram, but 3.2 has a y-axis limit to 2.5 to visualize the rest of the histogram and the KDEs better.
- From figure 3.1 we can see a staggering problem that the most jobs that are accessible within 1 hour by public transportation are nearing 0 with a frequency percentage of more than 17.5%,

outlining the clear-cut problem that most Egyptians don't have accessible commute methods to their jobs.

- Figure 3.2 outlines that the capital (Cairo) has the most spanned and diverse distribution of (jobs_perc_60), followed by Giza, Al-Qalyubia and lastly, Al-Sharkia

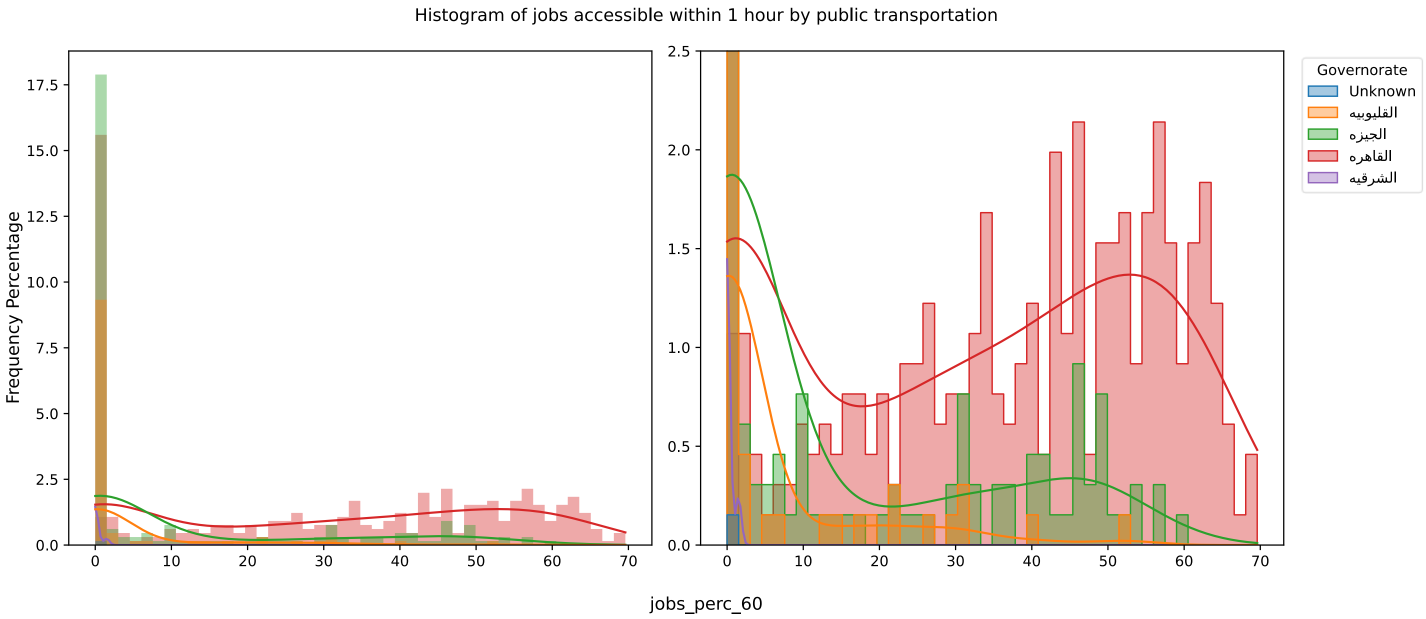


Figure 3.1 Histogram of (job_perc_60) & Figure 3.2 Histogram of (job_perc_60) with "2.5" y limit

Scatter plot:

- Figure 4 shows a scatterplot between population count and the number of total jobs, accompanied by the governates as a hue to the recorded points in the scatter plot, both axes are logarithmically scaled.
- It becomes apparent that there is a positive trend between the number of total jobs in an area and its apparent population count.
- The hue also represents that the highest number of jobs (around and more than 1 million) are concentrated around Cairo, regardless of the population count and the rest of the governates face a more diverse job count with a diverse population count.

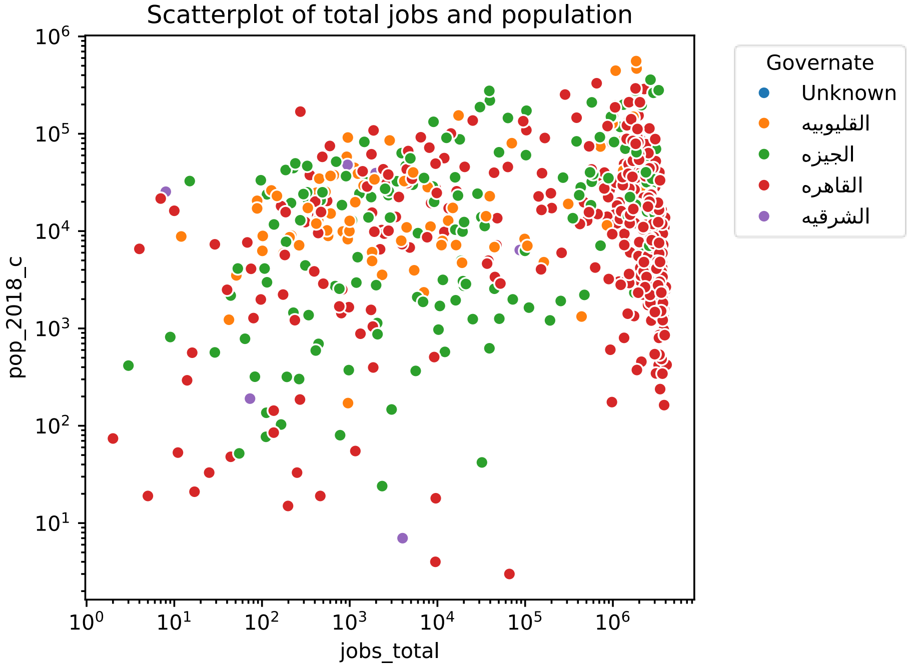


Figure 4. Scatter plot of (pop_2018_c) & (jobs_total)

Choropleth:

- Figure 5 shows a choropleth as a GIS representation generated from combining the shape files from the original dataset, the combination of the shape files of The Humanitarian Data Exchange for Level 3 and Level 4 administrative boundaries of Egypt, and the OSM data for the Nile, Metro Lines, Highways and Primary roads of Egypt.
- The choropleth shows that transportation availability is very correlated with the percentage of jobs accessible within 60 minutes, especially the metro lines highlighting the need for better transportation infrastructure to accommodate the population and giving people better access to job opportunities with reasonable commute time

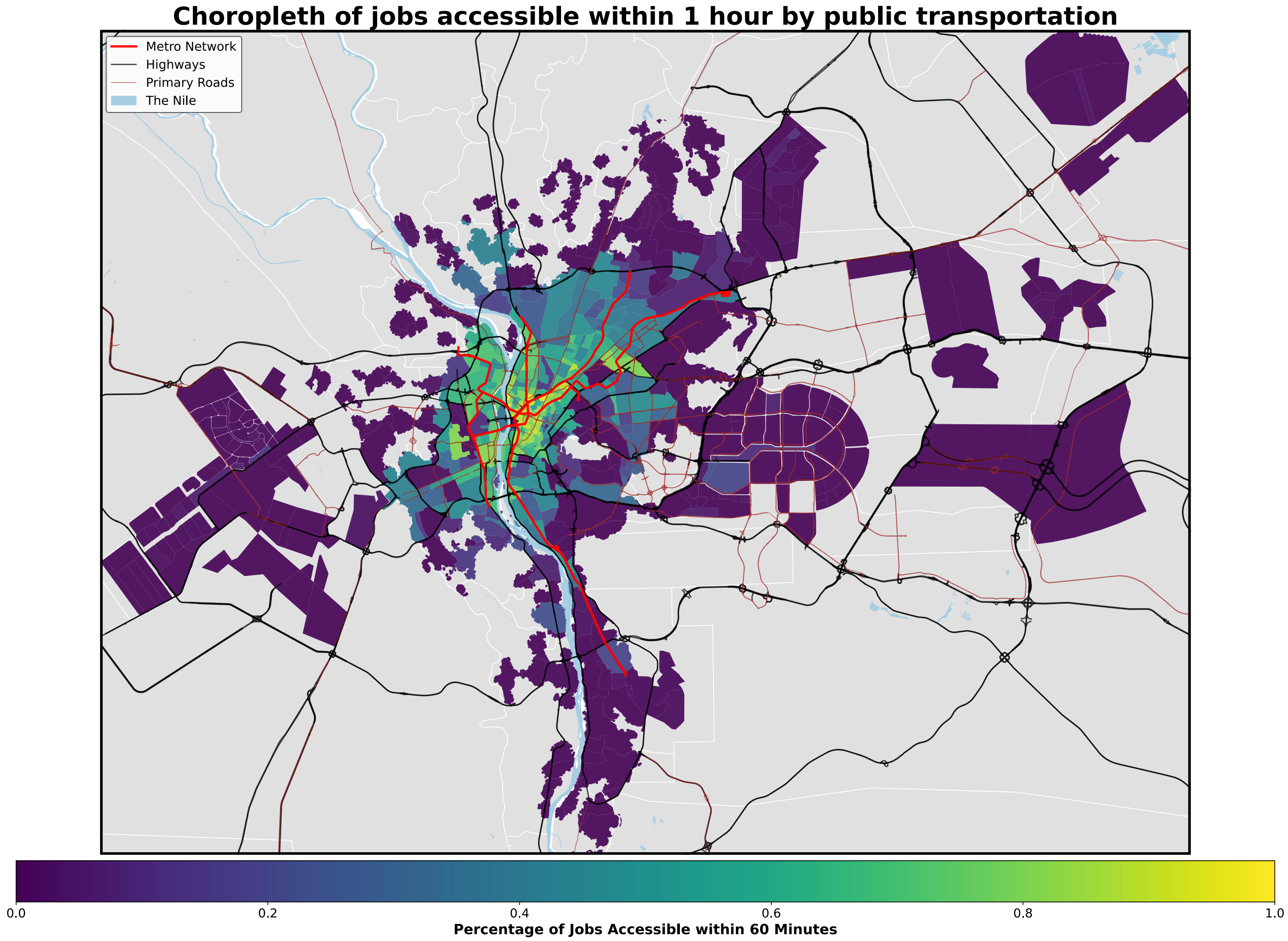

Choropleth of jobs accessible within 1 hour by public transportation
Metro Network
Highways
Primary Roads
The Nile
0.0
0.2
0.4
0.6
0.8
1.0
Percentage of Jobs Accessible within 60 Minutes

Figure 5. Choropleth of (ogc_fid) & (jobs_perc_60) on both level 3 & level 4 shape files of Egypt

Correlation Heat map:

- Figure 6 shows a correlation heatmap between all variables, the categorical variables were all encoded using an ordinal encoder and the missing white spaces indicate that correlation was impossible to calculate for the two variables in question to the position of the space
- The correlation heatmap shows very interesting relationships:
- For (jobs_total) & (jobs_perc_60) there is a strong inverse correlation with (zoning_tfc), and a weak albeit notable inverse correlation with (sec_name).
- Also, there is a very strong direct correlation with (pop_2018_c) and (jobs_total_60), proving further the findings of figure 4 that there is a positive trend between both of these variables

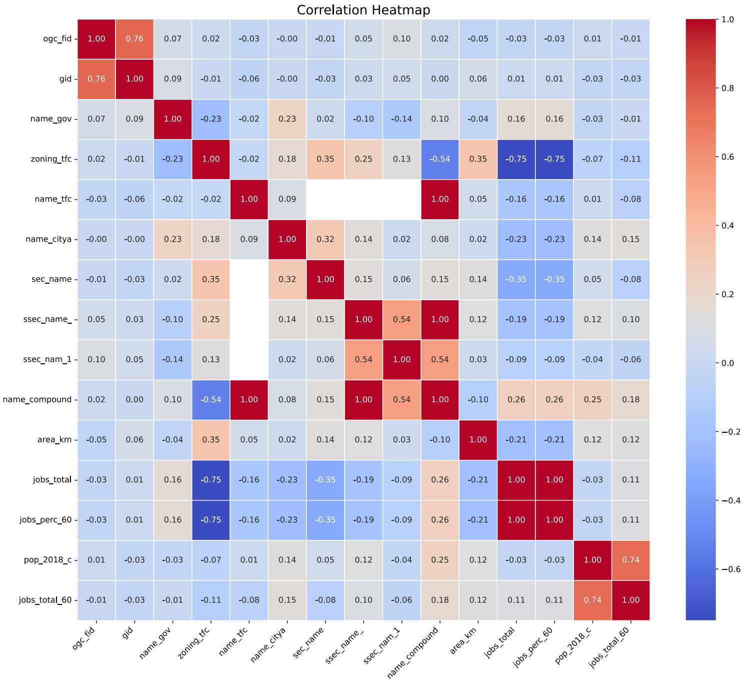


Figure 6. Correlation Heat map of all variables

Methodology:

1. Proposed Statistical Models:

Analysis of Variance (ANOVA):

- Research Question: RQ1 (Difference in accessibility by zoning).
- Relation to Findings: The Bar Plot (Figure 1) showed a massive visual gap where Central zones had >40% accessibility compared to <10% for others.

Hypothesis:

  - Ho (Null): The median job accessibility is equal across all zoning categories (Inner, Outer, Central).
  - Ha (Alternative): At least one zoning category has a significantly different median accessibility than the others.
  - Confidence Level: 95%

- Goal: To determine if zoning classification is a statistically significant determinant of transit equity, or if accessibility differences are merely due to random chance.
- Test Definition: We will perform a One-Way ANOVA test. We will treat (zoning_tfc) as the independent categorical factor (Levels: Inner, Outer, Central) and jobs_perc_60 as the dependent continuous variable. The Null Hypothesis is that the mean accessibility is equal across all three

zones. If the p-value is < 0.05, we reject the hypothesis and conclude that zoning plays a statistically significant role in transit equity.

Predictive Modelling (Logistic Regression):

- Research Question: RQ2 (Predicting Accessibility Status).
- Relation to Findings: The Histogram (Figure 3) showed a "zero-inflated" distribution, where districts largely fall into "Low Access" or "High Access" groups rather than a normal distribution.

Hypothesis:

  - Ho (Null): Demographic factors (Population and Total Jobs) have no effect on the likelihood of a district having "High Accessibility" (Coefficients $\beta = 0$).
  - Ha (Alternative): Population density and local job count are significant predictors of whether a district is well-connected to the transit network.
  - Confidence Level: 95%

- Goal: To build a binary classifier that quantifies how well simple demographic metrics can predict transit connectivity, thereby testing if "density equals access" in the Greater Cairo Region.
- Test Definition: We will categorize the target variable jobs_perc_60 into a binary outcome (0 = Low Access, 1 = High Access) using the median as a threshold. We will then train a Logistic Regression Classifier using (pop_2018_c) and (jobs_total) as features. We will evaluate the model's accuracy and the Odds Ratios to determine if demographic factors are reliable predictors of transit connectivity.

Pearson Correlation & Significance Test:

- Research Question: RQ3 (Population vs. Jobs Relationship).
- Relation to Findings: The Scatter Plot (Figure 4) indicated a positive linear trend.

Hypothesis:

  - Ho (Null): There is no linear correlation between Population and Total Jobs ($\rho = 0$).
  - Ha (Alternative): There is a non-zero linear correlation between where people live and where jobs are located.
  - Confidence Level: 95%

- Goal: To validate the urban planning assumption that population centers naturally attract employment hubs (agglomeration) and to quantify the strength of this relationship.
- Test Definition: We will calculate the Pearson Correlation Coefficient to quantify the strength of the linear relationship between Population and Total Jobs. To validate this result, we will perform a t-test for the significance of the correlation coefficient. This tests the Null Hypothesis that there is no association. A p-value < 0.05 will confirm that the positive trend observed in the scatter plot is statistically significant.

Pairwise F-Test for Equality of Variances:

- Research Question: RQ4 (Regional Economic Inequality across the Greater Cairo Region).
- Relation to Findings: The Box Plots (Figure 2) and descriptive statistics revealed a stark contrast in the spread of job opportunities between governorates. While Giza and Cairo displayed massive ranges (indicating high internal inequality between rich and poor districts), Sharkia and Qalyubia appeared compressed and uniform.

Hypothesis:

  - Ho (Null): The economic variance (inequality) is equal between the two governorates being compared ($\sigma_A^2 = \sigma_B^2$).
  - Ha (Alternative): One governorate has significantly higher economic variance than the other, indicating a more stratified economic structure.
  - Confidence Level: 95%

- Goal: To rigorously map the landscape of economic inequality by identifying specifically *which* governorates differ in their structural diversity (e.g., comparing stratified urban centers vs. uniform satellite zones).
- Test Definition: To rigorously map the landscape of economic inequality, we will perform Pairwise F-tests for Equality of Variances on the (jobs_total) variable. Instead of a single comparison, we will test all unique combinations of governorates (Cairo, Giza, Sharkia, Qalyubia).

2. Mathematical Models Used in the Analysis:

- Kruskal-Wallis H-Test:
  - Description: A non-parametric method used to test whether samples originate from the same distribution. In this study, it determines if there are statistically significant differences in the median employment access between the four governorates, serving as an alternative to ANOVA when normality assumptions are violated.
  - Variables:
    - N: Total number of observations in the dataset.
    - k: Number of groups (Governorates).
    - R_i: Sum of ranks for the i-th group.
    - n_i: Sample size of the i-th group.
  - Equation:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

- Mann-Whitney U:
  - Description: A non-parametric test used to compare differences between two independent groups. It is utilized here for post-hoc pairwise comparisons between specific governorates (e.g., Cairo vs. Giza) following a significant Kruskal-Wallis H-Test result.
  - Variables:
    - U: The test statistic.
    - n_1, n_2: Sample sizes of the two groups being compared.
    - R_1: Sum of ranks for the first group.
  - Equation:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

- Bonferroni Correction:
  - Description: A statistical adjustment applied to counteract the problem of multiple comparisons. It reduces the significance threshold (α) to minimize the risk of a Type I error (false positive) when conducting multiple pairwise Mann-Whitney U tests.
  - Variables:
    - $\alpha_{adjusted}$: The new, stricter significance threshold.
    - $\alpha_{original}$: The standard significance level (usually 0.05).
    - m: The number of hypotheses or pairwise comparisons being tested.
  - Equation:

$$\alpha_{adjusted} = \frac{\alpha_{original}}{m}$$

- Binary Logistic Regression:
  - Description: A statistical model used to predict the probability of a binary outcome. In this project, it estimates the probability of a district having "High Connectivity" based on demographic predictors like population and job density.
  - Variables:
    - P: The probability of the positive outcome (High Connectivity).
    - β_0: The intercept of the model.
    - β_1, β_2: The regression coefficients for the predictors.
    - X: The predictor variables (Population, Jobs).
  - Equation:

$$\ln\left(\frac{P}{1 - P}\right) = \beta_0 + \beta_1 X_{pop} + \beta_2 X_{jobs}$$

- Log-Log Pearson Correlation:
  - Description: A statistical measure used to quantify the strength and direction of the relationship between two variables that follow a power-law distribution. By applying a natural logarithmic transformation to both the population and employment data, the non-linear relationship is linearized, allowing the standard Pearson correlation formula to be accurately applied to the transformed values.
  - Variables:
    - r: The Pearson correlation coefficient for the log-transformed data.
    - n: The total number of data points (districts).
    - x_i': The natural logarithm of the independent variable (Population) for observation.
    - y_i': The natural logarithm of the dependent variable (Total Jobs) for observation.
    - $\bar{x}'$: The mean of the log-transformed independent variable.
    - $\bar{y}'$: The mean of the log-transformed dependent variable.
  - Equation:

$$r = \frac{\sum(x_i' - \overline{x'})(y_i' - \overline{y'})}{\sqrt{\sum(x_i' - \overline{x'})^2}\sqrt{\sum(y_i' - \overline{y'})^2}}$$

- Two-Sample F-Test:
  - Description: A parametric test used to assess whether two population variances are equal. It is applied here to determine if the economic inequality (variance in job distribution) in one governorate is significantly different from another.
  - Variables:
    - F: The test statistic.
    - S_{larger}^2: The larger sample variance.
    - $S_{smaller}^2$: The smaller sample variance.
  - Equation:

$$F = \frac{S_{larger}^2}{S_{smaller}^2}$$

- Ordinary Least Squares (OLS):
  - Description: A method for estimating the unknown parameters in a linear regression model. It operates by minimizing the sum of the squares of the vertical differences (residuals) between the observed responses and the predicted values.
  - Variables:
    - β: The parameter being estimated (slope or intercept).
    - y_i: The observed value for the i-th data point.

        - $\hat{y}_i$: The predicted value for the i-th data point.
    - Equation:

$$\min_{\beta} S(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Power-Law Scaling for Urban Statistics:
    - Description: A modelling approach used to capture the non-linear scaling relationships typical in urban systems. By applying a logarithmic transformation to both variables (Population and Jobs) before calculating the correlation, the power-law curve is linearized for analysis.
    - Variables:
        - r: The Pearson correlation coefficient of the transformed data.
        - x_i': The natural logarithm of the population ($\ln(P)$).
        - y_i': The natural logarithm of the job total ($\ln(J)$).
    - Equation:

$$J = \alpha P^{\beta} \xrightarrow{\ln} \ln(J) = \ln(\alpha) + \beta \ln(P)$$

- Classification Metrics:

Description: A set of metrics derived from the Confusion Matrix used to evaluate the effectiveness of the Logistic Regression classifier in correctly distinguishing between districts with high versus low connectivity.

Variables: TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative).

- Accuracy:
    - Description: The proportion of total predictions (both positive and negative) that were correct.
    - Equation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision:
    - Description: The proportion of positive identifications that were actually correct
    - Equation:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall:
    - Description: The proportion of actual positives that were identified correctly.

    - Equation:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score:
    - Description: The harmonic mean of precision and recall, providing a single metric that balances both concerns, particularly useful for uneven class distributions.
    - Equation:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Analysis Results:

RQ1 Analysis: Regional Differences in Employment Access:

1. Refuting The Use of ANOVA:

The problem with using ANOVA directly is that it requires multiple assumptions to be correct:

a. Normality:

Normality is that the dependent variable's distribution with each group should be normal.

Figure 7 shows a Q-Q (Quantile - Quantile) plot where, if the distribution is normal the ordered quantiles in y-axis would have a relationship with the theoretical quantiles of a normal distribution close to the reference red straight line.

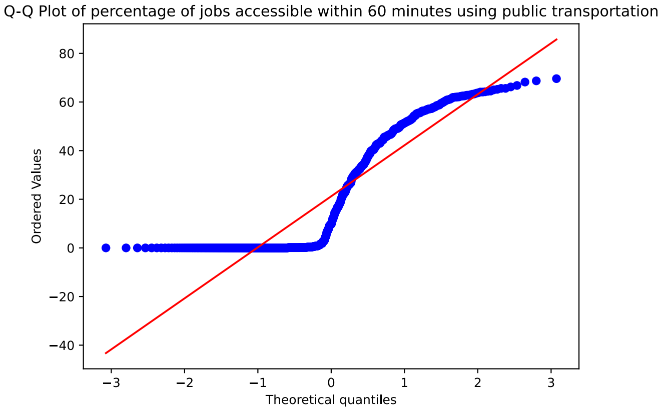


Figure 7. Q-Q plot of (jobs_perc_60)

As it appears that it doesn't follow the expected reference, it is not a normal distribution.

b. Homoscedasticity:

Homoscedasticity is that the variance of the dependent variable should be equal across all groups.

Figure 1 clearly shows that Homoscedasticity is not achieved due to the difference between the whiskers and the size of the boxes is not equal, refuting the need to use complex tests like Levene's test to verify further.

c. Independence:

Independence is that observations within and between groups must be independent, where one data point shouldn't influence another.

Figure 8 shows a residual plot accompanied by the Durbin-Watson value in the title, where the metric scales between 0 and 4 where a score of 2 indicates perfect independence.

The residual plot uses OLS (Ordinary Least Squares) to make a linear regression model, where it takes the residuals and plots them in a scatter plot indicates that if the points are scattered randomly, the residuals are independent, and if they form any pattern then the residuals are dependent.

The figure shows and the associated Durbin-Watson value indicate near perfect independence.

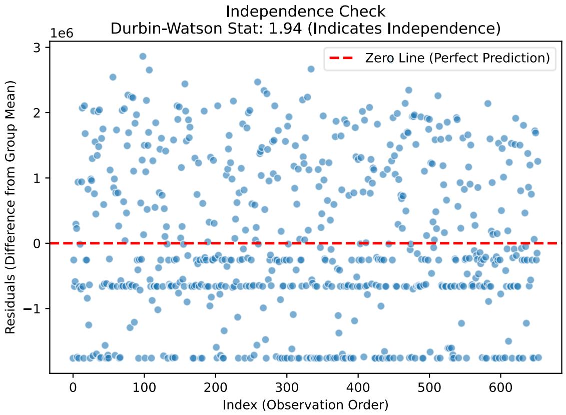
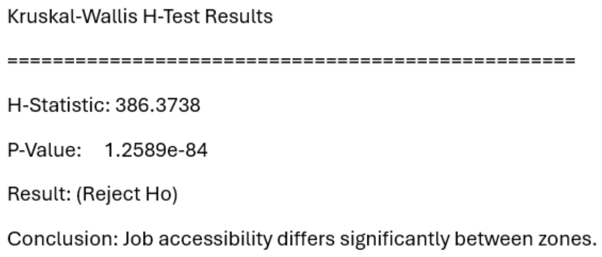


Figure 8. Residual Plot of (jobs_total) Grouped by (name_gov)

2. Using the alternative (Kruskal Wallis H-Test):

Since we cannot use ANOVA as it failed both Normality & Homoscedasticity requirements, we have to use an alternative test to compensate: Kruskal Wallis H-Test, where it is a non-parametric version of ANOVA.

Figure 9 shows the test results, and the H value has been converted to a p value, and due to its unimaginably small scale, we reject H_o and conclude that: Job Accessibility indeed differs significantly between zones, and it is not due to random implicit noise in the data.



```
Kruskal-Wallis H-Test Results
==================================================
H-Statistic: 386.3738
P-Value:    1.2589e-84
Result: (Reject Ho)
Conclusion: Job accessibility differs significantly between zones.
```

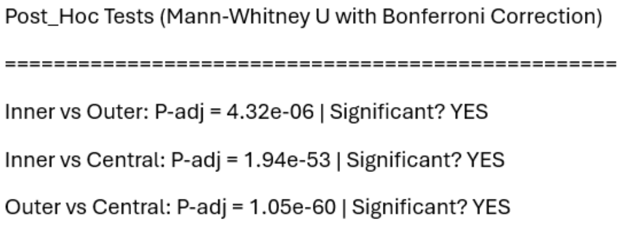
Figure 9. Kruskal Wallis H Test Results

3. Post H_{oc} Test:

Due to Kruskal Wallis H Test being concerned about a general check of a difference between median values among three or more independent groups, a post-H_{oc} test enables us to identify which groups caused the disruption in the proof (if happened), highlighting finer details that are crucial for clearer understanding.

A pairwise Mann-Whitney U test with Bonferroni correction was chosen and conducted as the post H_{oc} test.

Figure 10 shows the pairwise test results on all possible combinations (3) and the conclusion that there is a massive statistical difference in the means of every tested pair.



```
Post_Hoc Tests (Mann-Whitney U with Bonferroni Correction)
==================================================
Inner vs Outer: P-adj = 4.32e-06 | Significant? YES
Inner vs Central: P-adj = 1.94e-53 | Significant? YES
Outer vs Central: P-adj = 1.05e-60 | Significant? YES
```

Figure 10. Post Hoc Pairwise Test

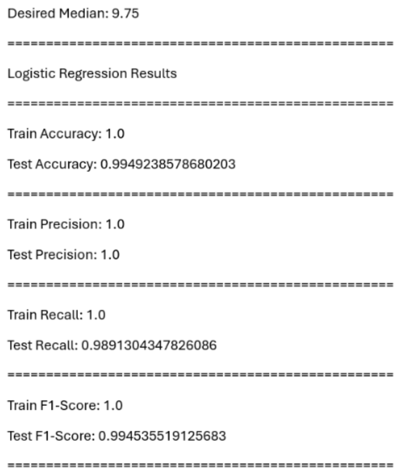
RQ2 Analysis: Predictive Modeling of Accessibility of Jobs:

A Logistic Regression Model has been used to see if population size (pop_2018_c) and local job count (jobs_total) are strong predictors for job availability percentage (job_perc_60).

(job_perc_60) has been binary encoded with a threshold equaling the median of that specific feature, due to the data being zero-inflated.

1. Test Scores:

Figure 11 shows the results of the model with different metrics, namely Accuracy, Precision, Recall, and F1-Score, segmented by train then test set scores, the result shows a stunning strong prediction ability and that proves the ability for both variables to be strong predictors.



```
Desired Median: 9.75
==================================================
Logistic Regression Results
==================================================
Train Accuracy: 1.0
Test Accuracy: 0.9949238578680203
==================================================
Train Precision: 1.0
Test Precision: 1.0
==================================================
Train Recall: 1.0
Test Recall: 0.9891304347826086
==================================================
Train F1-Score: 1.0
Test F1-Score: 0.994535519125683
==================================================
```

Figure 11. Threshold Value & Metrics of The Model

2. Test Set Confusion Matrix:

Figure 12 shows the confusion matrix on the test set (unseen data) showing further that aforementioned strong prediction ability.

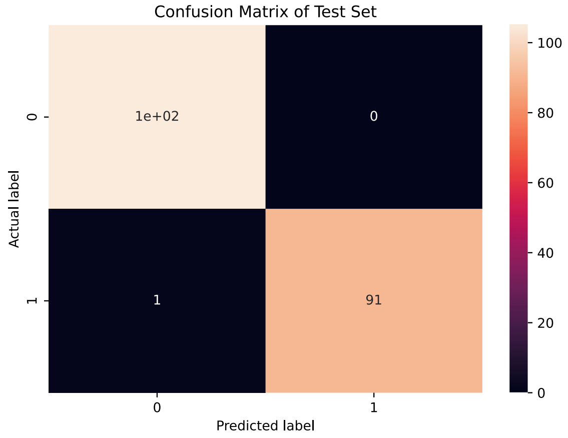


Figure 12. Confusion Matrix of The Test Set

3. Feature Importance:

Figure 13 shows the coefficients the machine learning algorithm assigned to each input feature, highlighting that the bigger the coefficient, the bigger its importance in the final classification decision.

We conclude that both (jobs_total) & (pop_2018_c) are extremely strong predictors of (jobs_perc_60).

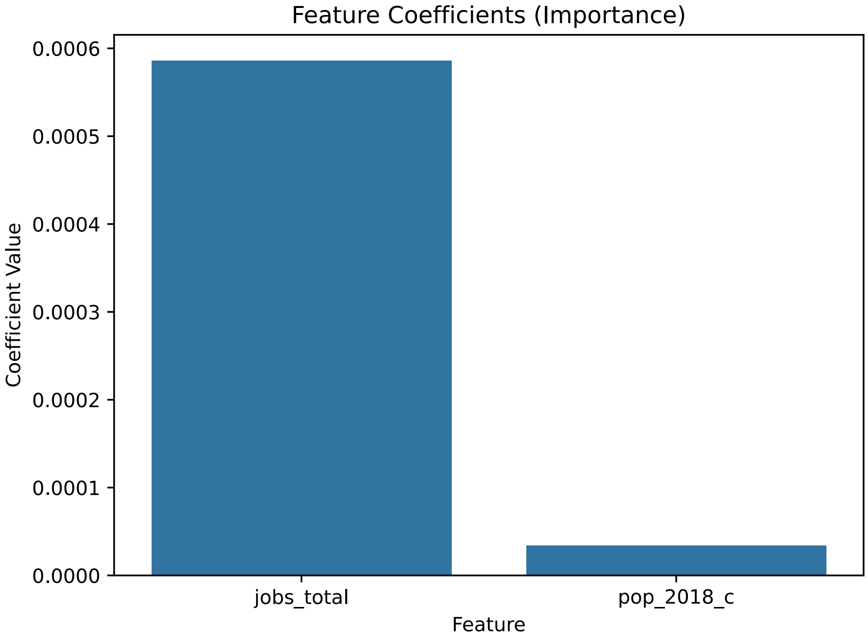
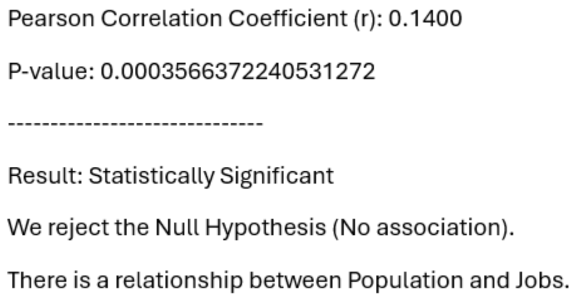


Figure 13. Feature Importance Bar plot

RQ3 Analysis: The Relationship Between Population and Employment:

A standard linear analysis was not possible due to urban scaling theory stating that population & jobs are related by a power law.

Fitting a linear relationship like Pearson's Correlation Coefficient is impossible so a compromise was made where we take a log-log transformation to enable it to become a linear relationship



```
Pearson Correlation Coefficient (r): 0.1400
P-value: 0.0003566372240531272
------------------------------
Result: Statistically Significant
We reject the Null Hypothesis (No association).
There is a relationship between Population and Jobs.
```

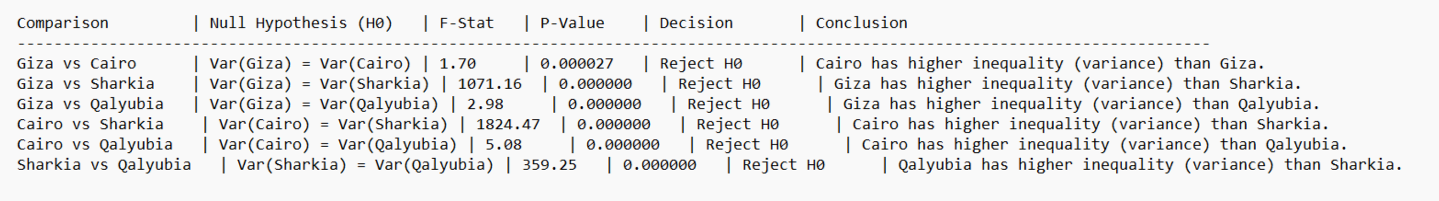
Figure 14. Log-Log Pearson Correlation Coefficient Test Results

Figure 14 shows the results of the analysis and confirms the findings noticed in figure 4 where there is indeed a relationship between (pop_2018_c) & (total_jobs).

RQ4 Analysis: Examining Regional Inequality (Variance):

A Pairwise F-Statistic Test was done as it was much more informative than Levene's Test which could only identify general Homoscedasticity without pinpointing the particular variances that caused the disruption, and as the number of possible combinations (6) was feasible enough for an exhaustive approach.

Figure 15 includes the test result of all possible pair combinations, and it shows that all H_0 Hypothesizes have been rejected, indicating massive economic regional inequality to all regions in the Greater Cairo Region (GCR).



```
Comparison         | Null Hypothesis (H0)   | F-Stat   | P-Value    | Decision       | Conclusion
--------------------------------------------------------------------------------------------------------------------------------
Giza vs Cairo      | Var(Giza) = Var(Cairo) | 1.70     | 0.000027   | Reject H0      | Cairo has higher inequality (variance) than Giza.
Giza vs Sharkia    | Var(Giza) = Var(Sharkia) | 1071.16  | 0.000000   | Reject H0      | Giza has higher inequality (variance) than Sharkia.
Giza vs Qalyubia   | Var(Giza) = Var(Qalyubia) | 2.98     | 0.000000   | Reject H0      | Giza has higher inequality (variance) than Qalyubia.
Cairo vs Sharkia    | Var(Cairo) = Var(Sharkia) | 1824.47  | 0.000000   | Reject H0      | Cairo has higher inequality (variance) than Sharkia.
Cairo vs Qalyubia   | Var(Cairo) = Var(Qalyubia) | 5.08     | 0.000000   | Reject H0      | Cairo has higher inequality (variance) than Qalyubia.
Sharkia vs Qalyubia   | Var(Sharkia) = Var(Qalyubia) | 359.25   | 0.000000   | Reject H0      | Qalyubia has higher inequality (variance) than Sharkia.
```

Figure 15. Pairwise F-Statistic Test Results

Conclusion:

The main findings are all pointing to the spatial mismatch problem that was aforementioned at the start of this document as all main findings state as follows:

Key Findings:

- RQ1 -> There is indeed a massive difference between the zones of a particular district and the accessibility of jobs in that same district.
- RQ2 -> Job Accessibility can be predicted effectively for each subsector's population & total amount of jobs, aiding in an initial prediction of any subsector that might have the problem.
- RQ3 -> In a subsector, the population number has a weak positive correlation with the number of jobs, which aids in weakening the general belief that a large population always

nets a large number of available job opportunities, which ties to the definition of spatial mismatch.

    - RQ4 -> There is also a large, real, and concerning variance between the number of available job opportunities and each governate, indicating major economic inequality where variances (inequality) of central governates (Cairo, Giza) are more prominent than other ones on the outskirts (Sharkia, Qalyubia), which are also one of the core root reasons od the spatial mismatch problem.

Future supporting research is strongly advised from different but strongly related datasets provided by Transport for Cairo Datasets Portal .

Complementary datasets to consider from:

- Transport for Cairo: Public Transport Vehicles Flow & Public Transport Passenger Flow: which are very complementary datasets that look into the vehicular and pedestrian traffic at morning peak time across the Greater Cairo Region (GCR).
- Transport for Cairo: Daily Passenger Boarding & Alighting by Public Transport Agency: this dataset is also extremely compatible because it shows the daily boarding and alighting activity of all possible stopping locations for public transportation across the Greater Cairo Region (GCR).

All paths this study approached and finished all lead to one decisive conclusion, Central Egypt is in dire need of direct and applicable action, without it spatial mismatch would get magnified to a point of no return.

References:

- Transport for Cairo: Population & Employment Access by District Dataset
- The Humanitarian Data Exchange: Egypt - Subnational Administrative Boundaries
- Geofabrik: OpenStreetMap for Egypt

Appendices:

Code for the Plots and Graphs (Made in a Python Juypter Notebook)