Data Acquisition, Analysis and Visualization

INFO8126 – Winter 2023 – Section 4 – Group 1

Final Group Project

Team Members:

Aesha Bhatt

Abidemi Adegbite

Raeshvanth Asokan

Rahul Malegavi

Table Of Contents

Summary:

Selecting population data offers valuable insights into demographic trends, urbanization, and regional dynamics, essential for various sectors like urban planning, healthcare, education, and economic development. Understanding population distribution, growth patterns, and demographic shifts helps policymakers, businesses, and communities make informed decisions regarding resource allocation, infrastructure development, and service provision. Additionally, population data serves as a key indicator of societal well-being and economic vitality, making it a crucial metric for monitoring and planning purposes in Canada (2014-2024).

Introduction:

This report analyzes population data from various regions of Canada between 2014 and 2024. It aims to provide insights into population trends and their implications for decision-making. By examining quarterly population data, the report seeks to uncover patterns, identify key drivers of population growth, and offer recommendations. The report aims to elucidate the dynamics of population changes and their implications for stakeholders through statistical analysis, including linear regression.

Methodology Overview:

To perform diverse analyses, we employ a range of analytical techniques. These include descriptive statistics, data visualization, descriptive data mining, statistical inference, linear regression, time series analysis and forecasting, predictive data mining, and spreadsheet models. Each team member specializes in a specific methodology to effectively uncover patterns, trends, and valuable insights.

Data collection: The procedure for collecting relevant and reliable information from a range of sources or using predetermined techniques.

Data cleaning: The process of finding and fixing errors, discrepancies, and missing values in the gathered data to guarantee its accuracy and dependability.

Descriptive analysis: The examination and summary of the most important traits, trends, and aspects of the data gathered to understand and comprehend its fundamental qualities.

Data visualization: Data is presented visually in forms like maps, graphs, and charts to help in understanding, examining, and sharing patterns, trends, and correlations found in the data.

Linear regression: A statistical technique that looks at the link between one or more independent variables and a dependent variable to forecast and comprehend the kind and strength of that relationship.

5 Techniques used for analysis:

Descriptive Statistics

Using graphical representations like histograms or bar charts, together with statistical measures like mean, median, mode, and standard deviation, descriptive statistics is the process of summarising and characterizing data. It offers a thorough summary of the data, allowing a more thorough understanding of its distribution, variability, and fundamental characteristics.

Population	
Mean	37407079.07
Standard Error	235900.2326
Median	37336956
Mode	#N/A
Standard Deviation	1510498.497
Sample Variance	2281605709195.22
Kurtosis	-0.611570154
Skewness	0.447855555
Range	5522867
Minimum	35247023
Maximum	40769890
Sum	1533690242
Count	41
Confidence Level(95.0%)	476772.1547

The given statistics provide information about a population. The population mean is calculated to be approximately 37,407,079.07. The standard error indicates the precision of the mean estimate, with a value of 235,900.2326. The median, which represents the middle value of the population, is found to be 37,336,956.

No mode is available in the data (indicated as #N/A), meaning no value appears more frequently than others. The standard deviation measures the variability of the population data and has a value of 1,510,498.497. The sample variance, a measure of how spread out the data is, is computed as 2,281,605,709,195.22.

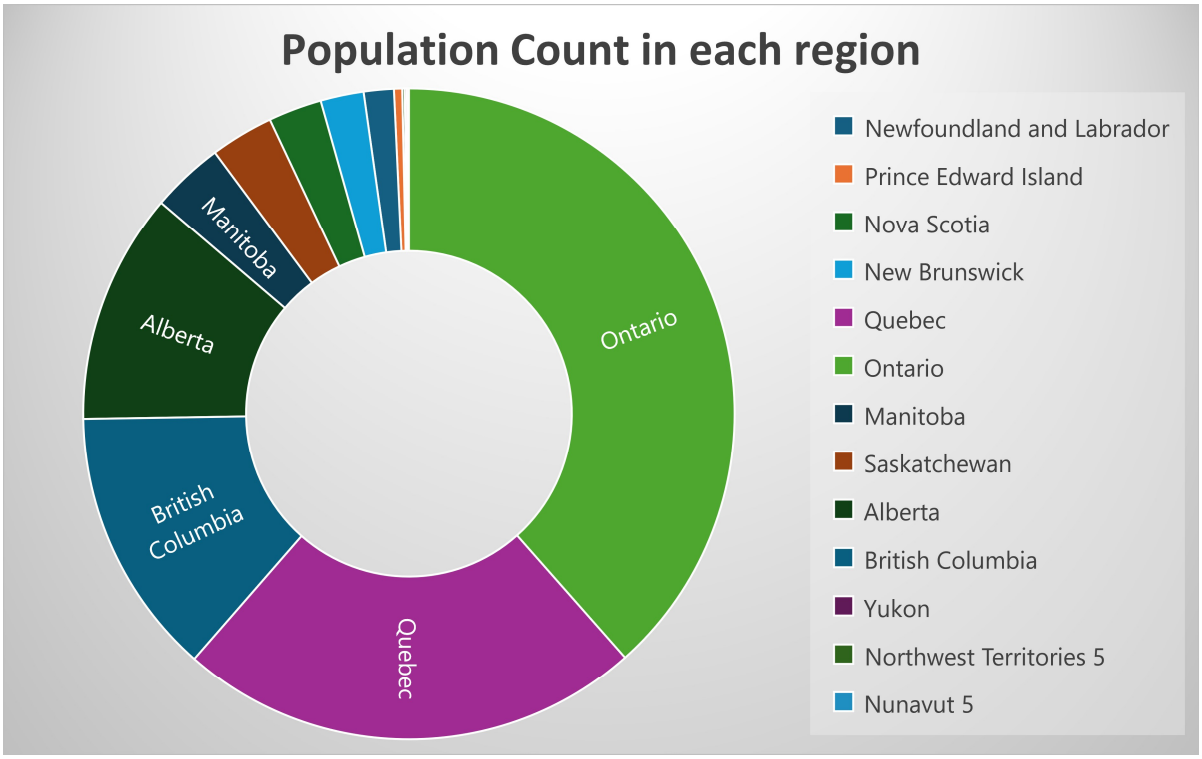
Kurtosis is a measure of the shape of the distribution, and a value of -0.611570154 suggests a relatively flat distribution compared to a normal distribution. Skewness measures the asymmetry of the distribution, and a positive value of 0.447855555 indicates a slight right-skewness.

The range of the data, which represents the difference between the maximum and minimum values, is 5,522,867. The minimum value in the population is 35,247,023, while the maximum is 40,769,890. The sum of all the values in the population is 1,533,690,242, and the count represents the number of observations in the population, which is 41. Lastly, the confidence level at 95.0% is provided as 476,772.1547, which indicates the margin of error for estimating population parameters based on the given sample data.

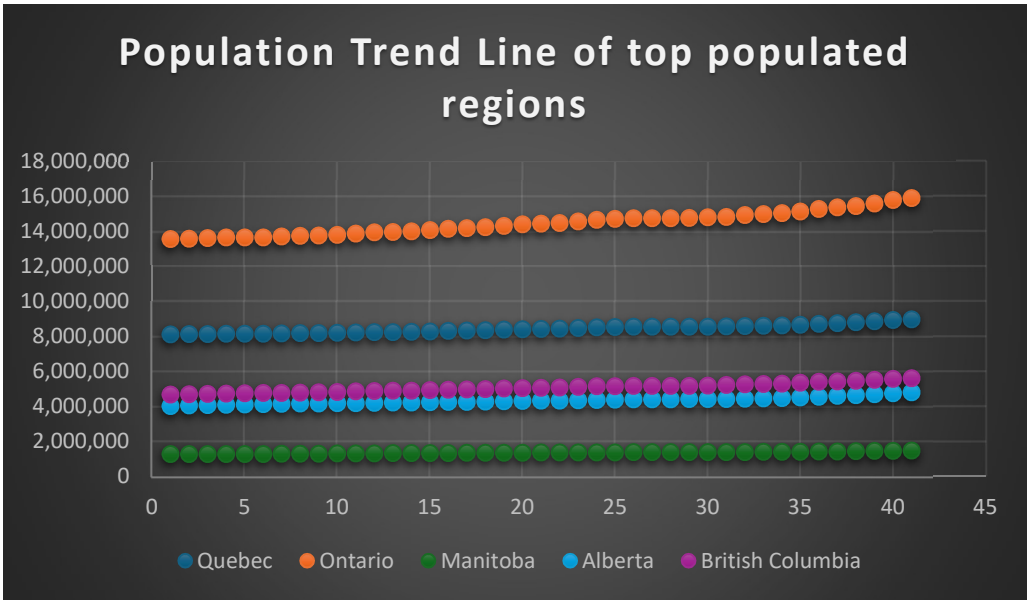
Data Visualization:

The technique of visually expressing data using graphs, charts, and other visual components is known as data visualization. Including a more intuitive and understandable representation of complicated data facilitates the identification of patterns, trends, and connections within the data.

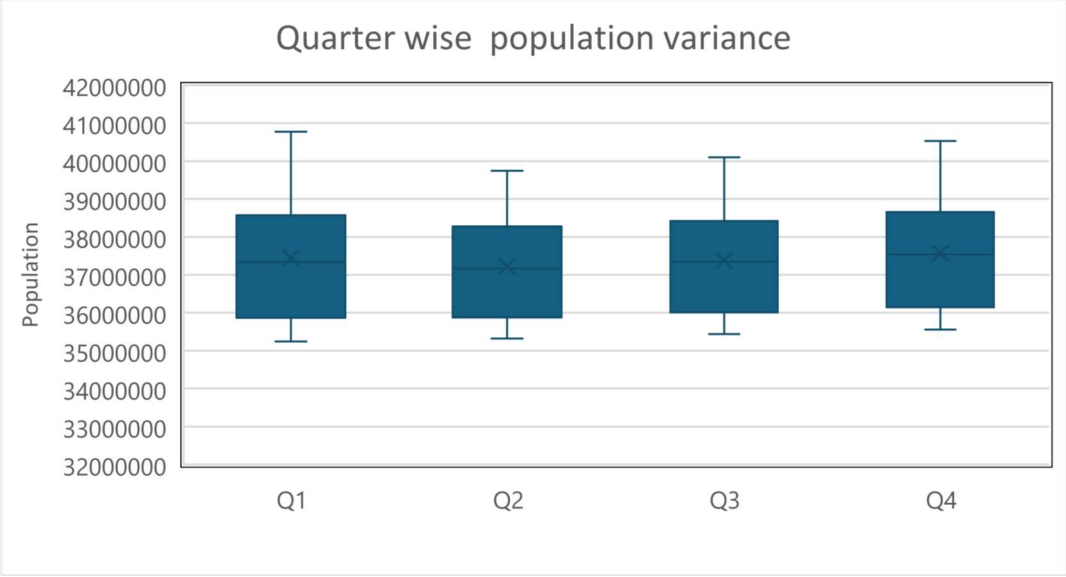
Analysis done:



Population Distribution: Ontario, Quebec, Manitoba, Alberta, and British Columbia are the most populous provinces in Canada. This could be attributed to various factors such as economic opportunities, urbanization, immigration patterns, and quality of life. These provinces may attract a larger influx of residents due to job opportunities, educational institutions, healthcare facilities, and cultural amenities.



Population Growth: The population in these provinces is growing steadily over time. This growth could be driven by natural increase (birth rates exceeding death rates) as well as net migration (more people moving into the provinces than leaving). Economic prosperity, infrastructure development, government policies, and social factors may contribute to the attractiveness of these provinces for individuals and families seeking better opportunities and quality of life.



Linear Regression:

A statistical method for modeling the connection between a dependent variable and multiple independent variables is called linear regression. The goal is to find the linear equation that best captures the relationship between the variables. We may examine the effects of independent factors on the dependent variable and provide model-based predictions using linear regression.

In linear regression, the dependent variable (Y) is the variable being predicted or explained, while the independent variable(s) (X) are the predictor variables. In the context of the population data, let's assume the dependent variable (Y) is the population, and the independent variables (X) are time (quarters) and geography.

$$Y\ (\text{Population}) = \beta 0 + \beta 1 X1 + \beta 2 X2 + \varepsilon$$

Where:

Y (Population) is the dependent variable

X1 is the independent variable representing time (quarters)

X2 is the independent variable representing geography.

β0 is the intercept (constant)

β1 and β2 are the coefficients associated with X1 and X2, respectively.

ε is the error term

The linear regression model estimates the coefficients (β0, β1, and β2) to minimize the sum of squared errors between the observed and predicted values of the population. The model then uses these coefficients to predict the population based on time (quarters) and geography values.

The formula provides a mathematical representation of the relationship between the population and the independent variables (time and geography), allowing for predictions of population changes over time and across different regions.

The linear regression output provides insights into the relationship between the variables in the population data. With an R-squared value of **0.9685**, it indicates that approximately **96.85%** of the variability in the population

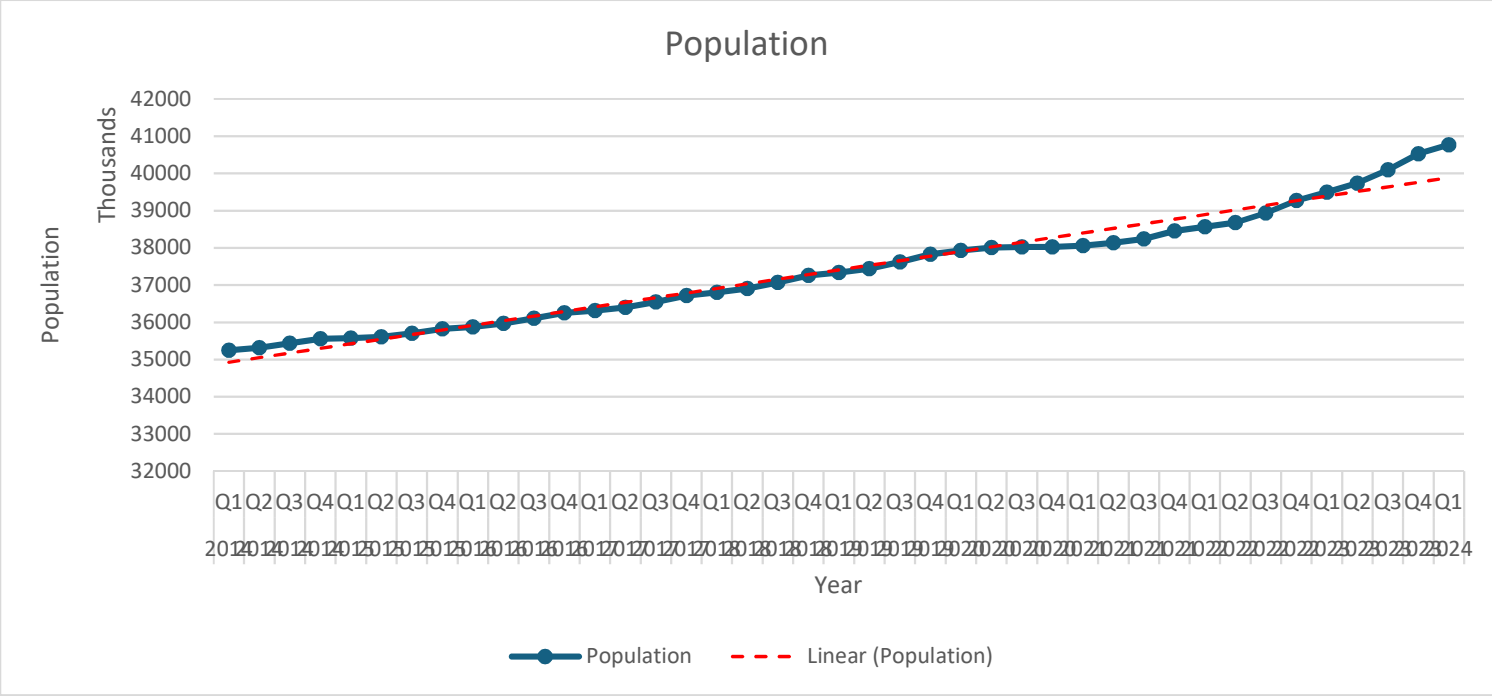
data can be explained by the independent variables considered in the model. **This suggests a strong linear relationship between the independent variables (geography and time) and the population.**

Time Series Analysis and Forecasting:

Time series analysis involves studying data collected over a period of time to identify patterns, trends, and seasonality. Forecasting, on the other hand, involves using historical data to make predictions about future values. By applying time series analysis and forecasting techniques, we can uncover insights about the past behavior of the data and make informed predictions about its future values.

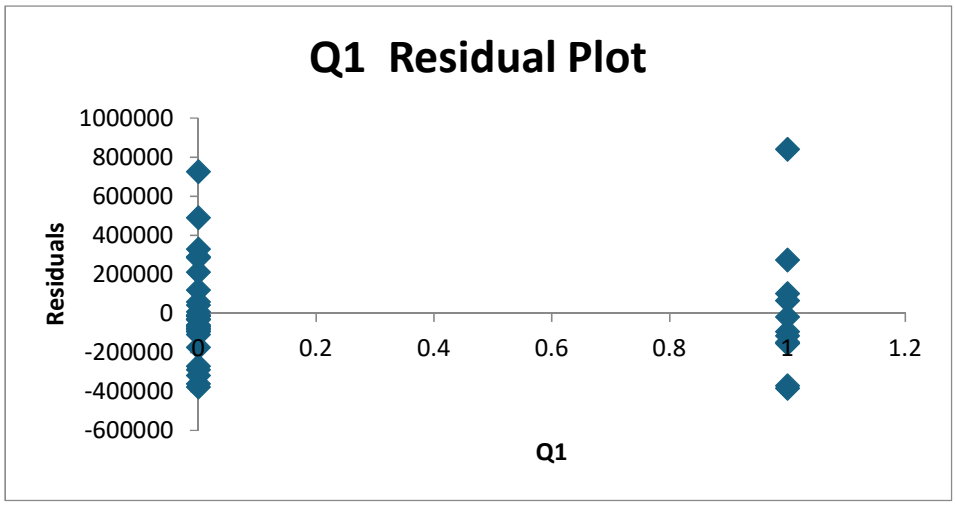
Analysis Done:

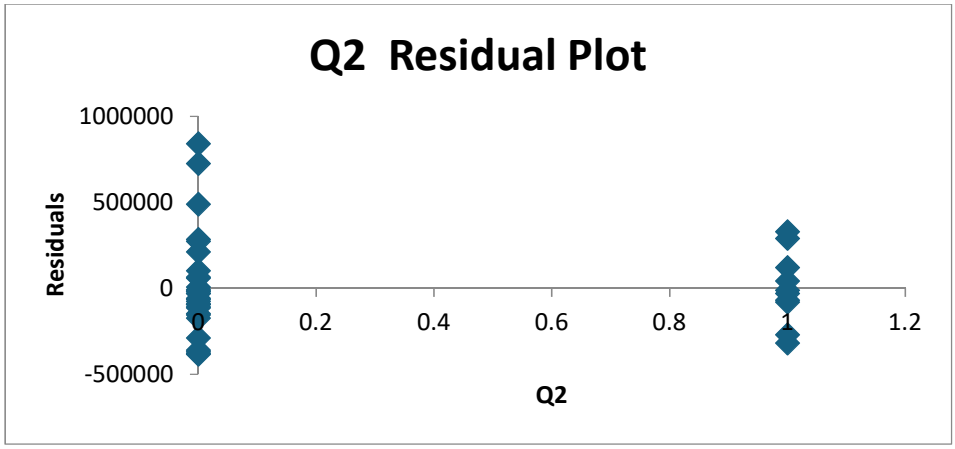
Year	Quarter	Population	Q1	Q2	Q3	Period
2014	1	35247023	1	0	0	1
2014	2	35320540	0	1	0	2
2014	3	35434066	0	0	1	3
2014	4	35555305	0	0	0	4
2015	1	35571043	1	0	0	5
2015	2	35606734	0	1	0	6
2015	3	35704498	0	0	1	7
2015	4	35823591	0	0	0	8
2016	1	35871484	1	0	0	9
2016	2	35970407	0	1	0	10
2016	3	36110803	0	0	1	11
2016	4	36257421	0	0	0	12
2017	1	36313068	1	0	0	13
2017	2	36397141	0	1	0	14
2017	3	36545075	0	0	1	15
2017	4	36722075	0	0	0	16
2018	1	36801579	1	0	0	17
2018	2	36903671	0	1	0	18
2018	3	37072620	0	0	1	19

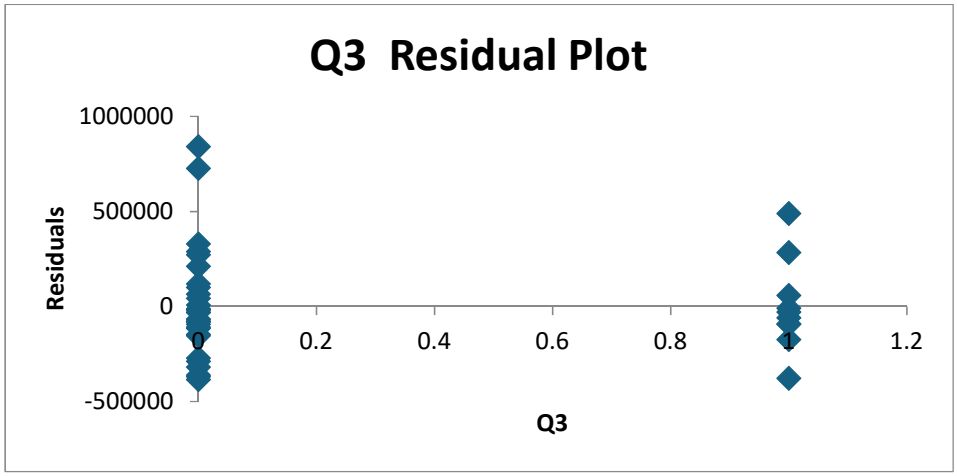


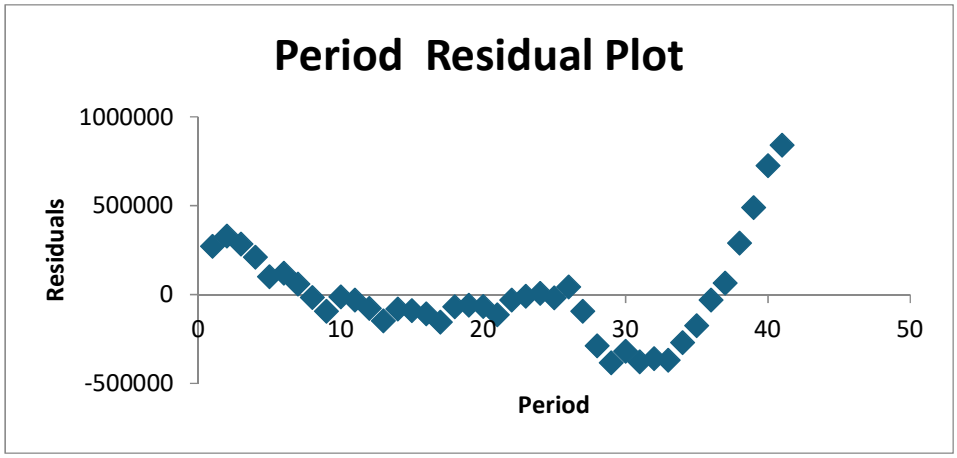
Using the previously mentioned statistics, we produced the Time-Series Chart below, which illustrates the population's positive skewness from 2014 to 2024.

Residual Plots:




Q2 Residual Plot
1000000
500000
0
-500000
Residuals
0
0.2
0.4
0.6
0.8
1
1.2
Q2


Q3 Residual Plot
1000000
500000
0
-500000
Residuals
0
0.2
0.4
0.6
0.8
1
1.2
Q3


Period Residual Plot
1000000
500000
0
-500000
Residuals
0
10
20
30
40
50
Period

Regression Statistics	
Multiple R	0.984145552
R Square	0.968542467
Adjusted R Square	0.965047186
Standard Error	282397.841
Observations	41

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	8.83933E+13	2.20983E+13	277.0999954	1.67748E-26
Residual	36	2.87095E+12	79748540575		
Total	40	9.12642E+13			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34848190.47	121337.5766	287.2003171	4.32753E-62	34602106.46	35094274.48	34602106.46	35094274.48
Q1	2103.092308	123444.9814	0.017036677	0.986501428	-248254.9339	252461.1186	-248254.9339	252461.1186
Q2	-104559.179	126512.7516	-0.826471464	0.413980995	-361138.9317	152020.5736	-361138.9317	152020.5736
Q3	-70003.48951	126347.3394	-0.554055905	0.582966143	-326247.7705	186240.7915	-326247.7705	186240.7915
Period	123852.4105	3733.905803	33.16966657	1.38824E-28	116279.6985	131425.1224	116279.6985	131425.1224

Year	Quarter	Prediction
2025	1	=34848190.47 + 2103.092308 * (A2 + (B2-1)/4) - 104559.179 * IF(B2=2,1,0) - 70003.48951 * IF(B2=3,1,0) + 123852.4105 * (A2-2014)
2025	2	=34848190.47 + 2103.092308 * (A3 + (B3-1)/4) - 104559.179 * IF(B3=2,1,0) - 70003.48951 * IF(B3=3,1,0) + 123852.4105 * (A3-2014)
2025	3	=34848190.47 + 2103.092308 * (A3 + (B3-1)/4) - 104559.179 * IF(B3=2,1,0) - 70003.48951 * IF(B3=3,1,0) + 123852.4105 * (A3-2014)
2025	4	=34848190.47 + 2103.092308 * (A5 + (B5-1)/4) - 104559.179 * IF(B5=2,1,0) - 70003.48951 * IF(B5=3,1,0) + 123852.4105 * (A5-2014)

Year	Quarter	Prediction
2025	1	39977394.32
2025	2	40039423.69
2025	3	40101453.05
2025	4	40163482.42

Residual Output:

RESIDUAL OUTPUT

Observation	Predicted Population	Residuals	Square
1	34974145.97	272877.028	74461872395
2	34991336.11	329203.8888	1.08375E+11
3	35149744.21	284321.7888	80838879593
4	35343600.11	211704.8888	44818959947
5	35469555.61	101487.386	10299689520
6	35486745.75	119988.2469	14397179383
7	35645153.85	59344.14685	3521727766
8	35839009.75	-15418.75315	237737948.6
9	35964965.26	-93481.25594	8738745213
10	35982155.4	-11748.3951	138024787.5
11	36140563.5	-29760.4951	885687068.9
12	36334419.4	-76998.3951	5928752849
13	36460374.9	-147306.8979	21699322170
14	36477565.04	-80424.03706	6468025738
15	36635973.14	-90898.13706	8262471322
16	36829829.04	-107754.0371	11610932503
17	36955784.54	-154205.5399	23779348524
18	36972974.68	-69303.67902	4802999926
19	37131382.78	-58762.77902	3453064198
20	37325238.68	-65753.67902	4323546305
21	37451194.18	-114238.1818	13050362185
22	37468384.32	-31141.32098	969781872.3

24	37820648.32	7513.679021		56455372.43
25	37946603.82	-18395.82378		338406332.4
26	37963793.96	43147.03706		1861666807
27	38122202.06	-93564.06294		8754233873
28	38316057.96	-288651.9629		83319955707
29	38442013.47	-383722.4657		1.47243E+11
30	38459203.6	-318285.6049		1.01306E+11
31	38617611.7	-377747.7049		1.42693E+11
32	38811467.6	-360013.6049		1.2961E+11
33	38937423.11	-369847.1077		1.36787E+11
34	38954613.25	-271046.2469		73466067933
35	39113021.35	-173965.3469		30263941906
36	39306877.25	-30737.24685		944778344.1
37	39432832.75	65185.25035		4249116863
38	39450022.89	289610.1112		83874016503
39	39608430.99	489330.0112		2.39444E+11
40	39802286.89	726109.1112		5.27234E+11
41	39928242.39	841647.6084		7.08371E+11
				2.87095E+12
			MSE=	70023108797

The Mean Squared Error (MSE) is a statistical measure used to assess the accuracy and performance of a regression or prediction model. In the provided assignment, the MSE value is reported as 70,023,108,797.

The MSE is calculated by taking the predicted values from the model and comparing them to the actual values. The differences between the predicted and actual values are squared to eliminate negative signs and emphasize larger errors. These squared differences are then averaged to obtain the MSE.

In this case, the high value of MSE indicates that, on average, the predicted values from the model have a substantial squared difference from the actual values. This suggests that the model's predictions may not align well with the true values and have a notable degree of error. The large MSE value implies the model's performance in accurately predicting the target variable is relatively poor.

It is important to note that a lower MSE value is desired, as it indicates a smaller average squared difference between the predictions and actual values. A lower MSE would suggest that the model's predictions are closer to the true values, indicating higher accuracy and precision.

In summary, the MSE value of 70,023,108,797 in this assignment indicates that the model's predictions may have a significant amount of error and may not accurately capture the true values of the target variable.

Statistical Inference:

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Avg 2014	13	35389233.5	2722248.731	1.65295E+13
Avg 2015	13	35676466.5	2744343.577	1.67563E+13
Avg 2016	13	36052528.75	2773271.442	1.71077E+13
Avg 2017	13	36494339.75	2807256.904	1.75694E+13
Avg 2018	13	37009338.75	2846872.212	1.81377E+13
Avg 2019	13	37555214	2888862.615	1.87377E+13
Avg 2020	13	37997798.25	2922907.558	1.9231E+13
Avg 2021	13	38222631.75	2940202.442	1.94529E+13
Avg 2022	13	38866584.75	2989737.288	2.01311E+13
Avg 2023	13	39965952	3074304	2.12832E+13
Avg 2024	13	40769890	3136145.385	2.22016E+13

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	2.33362E+12	10	2.33362E+11	0.012392605	0.999999992	1.903114608
Within Groups	2.48566E+15	132	1.88307E+13			
Total	2.48799E+15	142				

The ANOVA (Analysis of Variance) table you provided summarizes the results of the single-factor ANOVA test conducted on the population data across different years. Here's how to interpret the output:

Summary:

This section provides a summary of the data for each group (year) including the count (number of observations), sum, average (mean), and variance.

ANOVA:

Source of Variation: This column indicates the source of variation being analyzed. In this case, "Between Groups" refers to the variation between different years, while "Within Groups" refers to the variation within each year.

SS (Sum of Squares): This represents the sum of squares for each source of variation.

df (Degrees of Freedom): This indicates the degrees of freedom associated with each source of variation.

MS (Mean Square): This is the mean square for each source of variation, calculated by dividing the sum of squares by the degrees of freedom.

F (F-statistic): This is the F-statistic calculated by dividing the mean square for "Between Groups" by the mean square for "Within Groups."

P-value: This is the p-value associated with the F-statistic, indicating the probability of obtaining the observed F-value if the null hypothesis (no difference between group means) is true.

F crit (Critical F-value): This is the critical F-value corresponding to a chosen significance level and degrees of freedom.

Interpretation:

The F-statistic (0.012392605) is much smaller than the critical F-value (1.903114608) at the chosen significance level, and the p-value (0.999999992) is very close to 1.

We fail to reject the null hypothesis since the p-value is much larger than the significance level (typically 0.05). Therefore, there is no statistically significant difference in the population means across different years.

In other words, the average population across years does not vary significantly, and any observed differences are likely due to random variation rather than true differences between years.

Overall, the ANOVA test suggests that there is no significant difference in the average population across different years, based on the given data.

Further Analysis:

Seasonal Variations: The observation that the population tends to be high in the quarter of every year suggests a seasonal pattern. This could be related to various factors such as weather conditions, tourism, seasonal employment opportunities (e.g., agriculture, tourism, retail), academic calendars (e.g., the start of the school year), and migration patterns (e.g., seasonal workers, snowbirds).

Policy Implications: Understanding population trends and distribution across provinces can inform policymakers and stakeholders in areas such as urban planning, healthcare provision, education infrastructure, transportation, and social services. It may also guide decision-making related to immigration policies, economic development strategies, and resource allocation.

Future Planning: The consistent population growth in these provinces underscores the importance of long-term planning and investment in infrastructure, housing, healthcare, education, and other essential services to accommodate the needs of a growing population. Anticipating demographic changes and trends can help policymakers and organizations make informed decisions to support sustainable development and enhance the overall well-being of residents.

Conclusion:

Overall, analyzing population data provides valuable insights into demographic trends, regional dynamics, and socio-economic factors that shape the landscape of Canada's provinces. These insights can inform strategic planning, policy formulation, and resource allocation to address current challenges and capitalize on future opportunities for growth and development.

Insights from the population data reveal that Ontario, Quebec, Manitoba, Alberta, and British Columbia are the most populous provinces in Canada, experiencing steady growth over time. The population tends to peak each

quarter, suggesting seasonal influences like weather, tourism, and employment patterns. These trends inform policy decisions, urban planning, and resource allocation to accommodate growth and ensure sustainable development in key areas such as infrastructure, healthcare, and education.

RACI Chart:

Task/Activity	Rahul	Aesha	Raeshvanth	Abidemi
Descriptive Statistics	A	R	C	I
Data Visualization	I	C	R	A
Linear Regression	A	I	R	C
Time Series Analysis and Forecasting	A	R	C	I
Statistical Inference	R	I	A	C