

7COM1079-0901-2024

**Team Research and Development
Project**

**Final Report Title: Analysis of Russian
Demographic Data**

Group ID: A321

Dataset number: DS105 russian_demography.csv

Prepared by:

Monday Henry chukwu

Mani chandu kommaraneni

Muhammad Naveed

Ravikishore Kommina

github repo: <https://github.com/chukwuMonday/TeamResearch>

University of Hertfordshire

Hatfield, 2024

Table of Contents

1. Introduction.....	4
1.1 Problem statement and research motivation.....	4
1.2 The data set.....	4
1.3 Research question.....	4
1.4 Null hypothesis and alternative hypothesis (H0/H1).....	4
2. Background research.....	5
2.1 Research papers.....	5
2.2 Why RQ is of interest.....	5
3. Visualisation.....	6
3.1 Appropriate plot for the RQ <i>output of an R script</i>	6
3.2 Useful information for the data understanding.....	7
4. Analysis.....	7
4.1 Statistical test used to test the hypotheses and output.....	7
4.2 The null hypothesis is rejected /not rejected based on the p-value.....	8
5. Evaluation.....	8
5.1 What went well.....	8
5.2 Points for improvement.....	8
5.3 Group's time management.....	8
5.4 Project's overall judgement.....	9
6. Conclusions.....	9
6.1 Results explained.....	9
6.2 Interpretation of the results.....	9
6.3 Reasons and/or implications for future work, limitations of your study.....	9
7. Reference list.....	10
8. Appendices.....	11

8.1 R code used for analysis and visualization.....	11
---	----

List of Figures

Figure 1: Histograms of Birth rates of all the regions with normal curve	6
Figure 2: Region Distribution with Trend Ball	6
Figure 3: Box plot of Birth rate vs Region	7
Figure 4 Statistical Test	7

1. Introduction

1.1 Problem statement and research motivation

Russia has faced massive demographic changes, with birth rates oscillating in the regions, leading to concerns about population sustainability and regional disparities. Against this backdrop, policymakers need to understand these differences to design appropriate interventions that respond to declining birth rates and equitable resource distribution. Studies have identified regional inequalities in socio-economic development as a primary determinant of differences in birth rates (Shabunova and Rostovskaya, 2022). However, comprehensive statistical analyses comparing regional birth rates are scarce. The current study seeks to fill the gap by discussing regional variations that can guide evidence-based policy decisions addressing Russia's demographic challenges.

1.2 The data set

The dataset Russian Demography contains demographic and socio-economic data for regions of Russia across different years. The seven key variables include: year, representing the period region, indicating the specific geographic areas as npg, representing natural population growth is the birth rate and death rate, capturing vital demographic rates is gdw, representing gross domestic wealth, reflecting economic conditions and urbanization, representing the percentage of the urban population. This comprehensive dataset allows analysis of trends in demography, economic influence, and urbanization patterns across Russia.

1.3 Research question

The research question asks whether the mean birth rates across regions in Russia are significantly different. Regarding this, statistical analysis will be employed to analyse birth rates between regions depending on geographic characteristics, meaning that ANOVA or t-tests will smoothly perform to reveal regional differences in demographic rates.

1.4 Null hypothesis and alternative hypothesis (H0/H1)

The null hypothesis H0 is that there is no significant difference in the mean birth rates between the various Russian regions, meaning that the regional variations do not affect birth rates. The alternative hypothesis H1 is that there is a significant difference in the mean birth rates across regions, indicating that factors like socio-economic conditions or regional policies may influence these rates (Tkachenko, 2021). Testing of the above hypotheses will

be done using statistical tools like ANOVA, which will allow for the comparison of the mean birth rates between the regions and help in concluding whether any observed differences are statistically significant.

2. Background research

2.1 Research papers

The dataset on Russian demography has been useful in examining demographic patterns, socio-economic factors, and their impacts. Various other researchers employ identical data sets within the examination of demographic concerns. According to Pomazkin and Filippov, (2022), the first research paper titled ‘Demographic Resource for Data Analysis and Visualization’ described the population data and population dynamics. With birth and death rates, but more importantly the dynamics of figures that relate to urban status, its main point underlines the need for mapping and visualizing such data for pattern analysis and forecasting. This paper also demonstrates how datasets such as the Russian Demography dataset are useful in identifying regional disparities, and the socio-economic factors affecting demographic patterns. The second research paper of Pant, (2023), "Russia's Demographic Trajectory: 'Economic Development and Demographics': Key Dimensions and Implications" an article takes a close look at Russia's economic development amidst demographic changes. It goes into the details of the ongoing differentials in birth and death rates and associates them with things like; urbanization and economic inequality. This study therefore brings to light the need to approach policy issues and questions of sustainable development by engaging with demographic data. According to Simagin, (2021), the ‘Demographic Issues in Today's Russia: Natural Population Increase and Regional Disparities’ examines the demographic concerns of modern Russia, including natural population increment and disparities between the Russian regions. From this study, it is clear how demographic data enable one to track important trends like low birth rates and issues to do with urbanization among others that must inform policymaking.

2.2 Why RQ is of interest

This study addresses a critical gap in understanding regional disparities in birth rates across Russia. The literature already points to demographic challenges like falling birth rates and urbanization. Few studies focus on regional comparisons of mean birth rates using statistical techniques. This is significant because regional variations can provide insight into

socio-economic and policy-driven factors influencing demographics. This end would guide focused interventions and future research into addressing regional disparities. Examining these differences will further the understanding of Russia's demographic trends and support region-specific strategies for sustainable population growth and development.

3. Visualisation

3.1 Appropriate plot for the RQ output of an R script

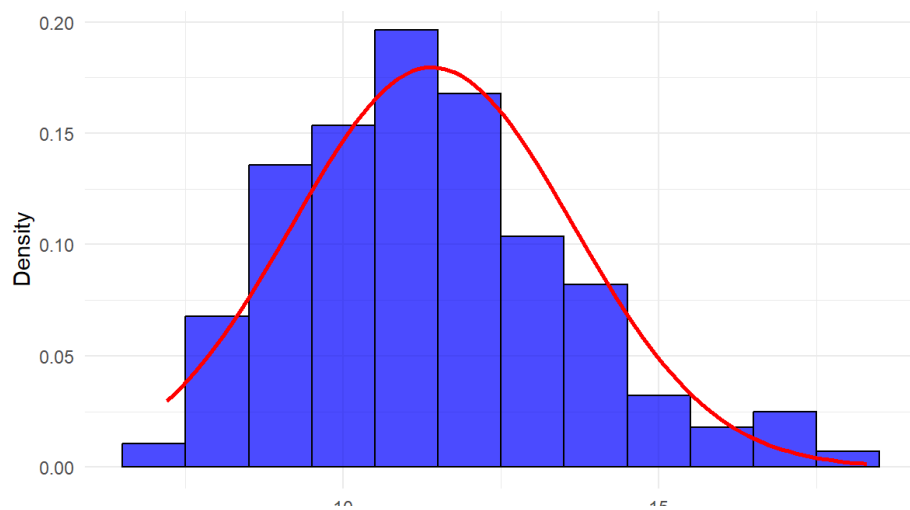


Figure 1: Histograms of Birth rates of all the regions with normal curve

The histogram shows a normal curve for the overall distribution of birth rates across all regions. It reflects a slightly skewed pattern. Such a visualization is chosen to watch the overall pattern, detect outliers, and see if birth rates approximate any normal distribution.

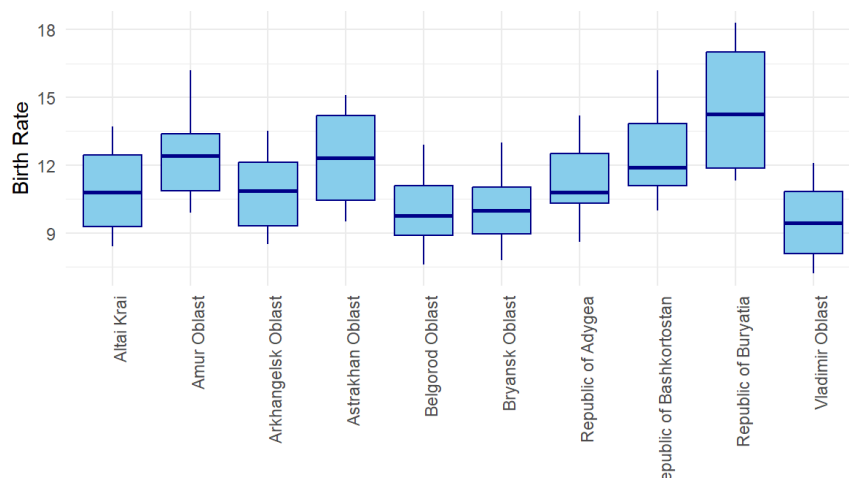


Figure 3: Box plot of Birth rate vs Region

The above box plot shows that there are about 10 to 30 various regions, showing variation in distribution. The plot exhibits some regions having higher variability and outliers; these differences show varying birth rates. The highest birth rate was observed in “Yemen Rep” and the lowest in “Bosnia & Herzegovina”.

3.2 Useful information for the data understanding

Histograms show that birth rates are skewed across all regions, which implies some regional outliers. The bar chart with a trend line indicates a variable trend in birth rates across regions. The box plot shows a large variability with a few regions showing higher disparities and outliers in birth rates.

4. Analysis

4.1 Statistical test used to test the hypotheses and output

```
> # Print the summary of the ANOVA result
> print(anova_summary)
              Df Sum Sq Mean Sq F value Pr(>F)
region          9  589.8   65.54   22.54 <2e-16 ***
Residuals     270  785.1    2.91
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # Extract p-value from the ANOVA summary
```

Figure 4 Statistical Test

The statistical test employed is the one-way ANOVA (Analysis of Variance). The appropriate test is because the research question (RQ) is to determine the mean birth rate across many regions, and the birth rate variable is continuous (Cheremisina and Cheremisina, 2024). The summary output gives degrees of freedom (Df), sum of squares (Sum Sq), mean square (Mean Sq), F value and probability value for the models. The region has 84 degrees of freedom, a sum of squares of 589.9, and a mean square of 65.54 specifically. There are 2.91 degrees of freedom, 7851.1 sums of squares and a mean square of 2.91 for the residuals. A highly significant difference in birth rates is shown across regions by an F value of 65.4 with a probability value less than 2e-16.

4.2 The null hypothesis is rejected /not rejected based on the p-value

The group rejects our null hypothesis, that different regions have the same birth rate, at a p value less than 2e-16. This really very small p-value implies that it is in fact nearly

impossible to observe the observed differences in birth rates due to random luck. Consequently, it is concluded that the mean birth rates for the different regions of Russia vary statistically significantly. This rejection adds momentum to the alternative hypothesis that regional birth rates are indeed quite different.

5. Evaluation

5.1 What went well

Overall, our group worked well together and efficiently divided tasks and worked together for data analysis, interpretation and writing. Group members did not take too much time to understand the dataset and perform the required statistical tests. Clear communication and assignment of labour divided the work among team members for the most efficient project.

5.2 Points for improvement

Areas to improve upon would include better initial planning and coordination of deadlines on each task on hand. Sometimes, work roles overlapped a bit and could have been more clearly defined. Besides, the group could have done more exploratory data analysis before members got to test the hypotheses.

5.3 Group's time management

All the members managed our time well, however there were bounces at first. For research, data analysis, and writing enough time is given by each member. But last-minute changes wound up piling pressure, whereas a more explicitly laid out timeline with internal checkpoints may have spared some rushed finishing work.

5.4 Project's overall judgement

Overall, the project was able to analyze and interpret Russian demographic data. The ANOVA test showed regional birth rate differences clearly. Nevertheless, for more in-depth exploratory analysis and smoother coordination, the group might have been able to reach further conclusions. However, the final resulting output was perfectly met and executed on the intended output.

6. Conclusions

6.1 Results explained

The ANOVA test results indicate a significant difference in birth rates in various parts of Russia. A value of the F (F statistic) of 65.54, and a p-value less than $2e-16$ suggest that the between-region variation in birth rates is not due to chance, but rather reflects real differences. This is evidence of the regional factors affecting the childbirth rates in Russia.

6.2 Interpretation of the results

The differences in birth rates between Russian regions are so big that regional factors like economic, cultural, and healthcare factors all appear to matter in demographic terms. This raises the possibility of targeted interventions for policymakers to close regional disparities in birth rates in order to affect population growth and the economy at a national scale.

6.3 Reasons and/or implications for future work, limitations of your study

Further work to expand these findings is possible by adding in other variables, such as economic status or access to healthcare, which may allow for a better explanation of regional variance. Such limitations arise from confining this analysis to a dataset that does not take into account other demographic phenomena (mortality and migration), thus giving a more complete picture of population dynamics.

7. Reference list

Cheremisina, N.V. and Cheremisina, T.N. (2024). Economic and statistical analysis of fertility in the region. *Scientific Works of the Free Economic Society of Russia*, [online] 249(5), pp.375–398. doi:<https://doi.org/10.38197/2072-2060-2024-249-5-375-398>.

Pant, H. (2023). *Russia's Demographic trajectory: dimensions and implications*. [online] orfonline.org. Available at: <https://www.orfonline.org/research/russia-s-demographic-trajectory-dimensions-and-implications> [Accessed 7 Jan. 2025].

Pomazkin, D. and Filippov, V. (2022). Demographic resource for data analysis and visualization. *Population and Economics*, [online] 6(3), pp.117–124. doi:<https://doi.org/10.3897/popecon.6.e81027>.

Shabunova, A.A. and Rostovskaya, T.K. (2022). Demographic Policy in Modern Russia: Population View and Expert Assessment. *Herald of the Russian Academy of Sciences*, [online] 92(6), pp.702–712. doi:<https://doi.org/10.1134/s1019331622050045>.

Simagin, Y. (2021). Results of the study of demographic problems in Russia in the 21st century. *POPULATION*, [online] 24(4), pp.4–22. doi:<https://doi.org/10.19181/population.2021.24.4.1>.

Tkachenko, A.A. (2021). SOCIO-ECONOMIC ASSESSMENT OF THE DEMOGRAPHIC SITUATION IN RUSSIA. *social & labor researches*, [online] 45(4), pp.89–97. doi:<https://doi.org/10.34022/2658-3712-2021-45-4-89-97>.

8. Appendices

8.1 R code used for analysis and visualization

```
# Load necessary libraries
library(dplyr)
library(ggplot2)
library(tidyr)

# Load the data
data <- read.csv("C:/Users/Henry/Documents/Research/TeamResearch/russian_demography.csv")

# Check for missing values
print(colSums(is.na(data)))

# Ensure 'region' column exists
if(!("region" %in% colnames(data))) {
  stop("The dataset does not contain the 'region' column.")
}

# Select 10 unique regions
selected_regions <- unique(data$region)[1:10] # Modify as needed to choose specific
regions
print(paste("Selected regions:", paste(selected_regions, collapse = ", ")))

# Subset the data to include only the selected regions
data_10_regions <- data %>%
  filter(region %in% selected_regions)

# Check for missing values in the subset
print(colSums(is.na(data_10_regions)))

# Ensure there are no missing values in 'birth_rate' or 'region'
```

```

data_clean_10_regions <- data_10_regions %>%
  drop_na(birth_rate, region)

# Perform one-way ANOVA to test the difference in birth rates between the 10 regions
anova_result <- aov(birth_rate ~ region, data = data_clean_10_regions)
anova_summary <- summary(anova_result)

# Print the summary of the ANOVA result
print(anova_summary)

# Extract p-value from the ANOVA summary
p_value <- anova_summary[[1]][["Pr(>F)"]][1]

# Interpret the p-value
if (!is.na(p_value) && p_value < 0.05) {
  print("There is a significant difference in the mean birth rates between the selected
regions.")
} else {
  print("There is no significant difference in the mean birth rates between the selected
regions.")
}

# Create a boxplot of birth rates by the selected regions
ggplot(data_clean_10_regions, aes(x = region, y = birth_rate)) +
  geom_boxplot(fill = "skyblue", color = "darkblue", outlier.color = "red", outlier.shape = 16)
+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Boxplot of Birth Rates by Selected Regions", x = "Region", y = "Birth Rate")

# Create a histogram of birth rates
ggplot(data_clean_10_regions, aes(x = birth_rate)) +

```

```
geom_histogram(aes(y = ..density..), binwidth = 1, fill = "blue", color = "black", alpha =
0.7) +
stat_function(fun = dnorm,
              args = list(mean = mean(data_clean_10_regions$birth_rate, na.rm = TRUE),
                           sd = sd(data_clean_10_regions$birth_rate, na.rm = TRUE)),
              color = "red", size = 1) +
labs(title = "Histogram of Birth Rates for Selected Regions with Normal Curve",
      x = "Birth Rate",
      y = "Density") +
theme_minimal()
```

Git commits

1. Commit Message: [introduction push](#)] updating the introduction and scope of our research
2. Commit Message: [cleanings](#)] pushed the r code for clearing datasets
3. Commit Message: [[final report upload](#)] uploaded final report on the statistical analysis carried out.

Git log output

commit 6f3eed087aa5a6c9afaf3d06b7f09f0fd214a59a

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Tue Jan 7 13:22:50 2025 +0000

cleaning data

Output/Rplot.png | Bin 6223 -> 0 bytes

Output/analysis 1 output.jpeg | Bin 269226 -> 0 bytes

test.R | 21 -----

3 files changed, 21 deletions(-)

commit f502f5f195291df8dee71542c0a70127fb51e926

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Tue Jan 7 13:14:38 2025 +0000

few touches

histogram upd.png | Bin 17582 -> 0 bytes

1 file changed, 0 insertions(+), 0 deletions(-)

commit ea99738a698bd625abfcd215fdc26cad85a6a51

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Jan 6 21:37:29 2025 +0000

new analysis

Analysis	of	test.R		67
+++++				
Analysis.R		50	+++++	-----
boxplot updated.png Bin 0 -> 18677 bytes				
histogram upd.png Bin 0 -> 17582 bytes				
histogram updated.png Bin 0 -> 17582 bytes				
5 files changed, 104 insertions(+), 13 deletions(-)				

commit 1739288265935e4e9d65ec2abd05b5b2d3de1f6b

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Jan 6 19:34:49 2025 +0000

rebases

Analysis.zip | Bin 0 -> 544627 bytes
Final Research Report.docx | Bin 64802 -> 0 bytes
GROUP A321 (1).pptx | Bin 152942 -> 0 bytes
GROUP A321.pptx | Bin 0 -> 209311 bytes

Introduction.txt | 3 +-
README.md | 93 ++++++-----
TeamResearch.Rproj | 4 +-
final ppt A321.pptx | Bin 0 -> 810386 bytes
test.R | 21 ++++++
9 files changed, 46 insertions(+), 75 deletions(-)

commit 1f9b0e7b95fec9917c8008617765b7f0359df7b3

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Jan 6 17:46:55 2025 +0000

final report upload

Final Research Report.docx | Bin 0 -> 64802 bytes
1 file changed, 0 insertions(+), 0 deletions(-)

commit 86f09593d677575159c176f395950baf8c148532

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Jan 6 17:30:48 2025 +0000

minor fixes 2

README.md | 10 +++-----
1 file changed, 3 insertions(+), 7 deletions(-)

commit 7b43360d6e9a5a14c9c48dd043e4f47825f9b218

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Jan 6 17:23:17 2025 +0000

fix for README.md

README.md | 24 ++++++-----
1 file changed, 22 insertions(+), 2 deletions(-)

commit 0617c055f0bd7f2e8f538b3b04f6bec43a83398c

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Jan 6 17:18:56 2025 +0000

repo cleaning

GROUP A321.pptx | Bin 209311 -> 0 bytes

Untitled | 1 -

final ppt A321.pptx | Bin 810386 -> 0 bytes

test.R | 21 -----

4 files changed, 22 deletions(-)

commit 44275157a11124ec73261bd41d4c569878ba70bb

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Jan 6 17:13:05 2025 +0000

typo fix

README.md | 2 +-

1 file changed, 1 insertion(+), 1 deletion(-)

commit 8aa3eb744ef0b5daf80f3efd577f4e34d36a5136

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Jan 6 17:10:06 2025 +0000

minor fixs

GROUP A321 (1).pptx | Bin 0 -> 152942 bytes

Introduction.txt | 3 +-

README.md | 77 ++++++-----

3 files changed, 56 insertions(+), 24 deletions(-)

commit 4f4affaae49f03f2d04e6fa08b0149f57f3d23ba

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Jan 6 15:27:05 2025 +0000

testing acc

TeamResearch.Rproj | 3 +++

Untitled | 1 +

2 files changed, 4 insertions(+)

commit c97a77b4f74fa2dd9c19d890d3e381570e3f8681

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Mon Nov 25 08:49:21 2024 +0000

cleanings

Analysis.R | 115 ++++++++-----

Output/Rplot.png | Bin 9294 -> 6223 bytes

.../analysis 1 output.jpeg | Bin

Output/bell curve.png | Bin 5488 -> 0 bytes

Output/box plot.png | Bin 50648 -> 0 bytes

analysis 1.R | 40 -----

analysis 2 output.jpeg | Bin 77350 -> 0 bytes

analysis 2.R | 49 -----

analysis 3 output.jpeg | Bin 56221 -> 0 bytes

analysis 3.R | 17 ---

analysis 4 output.jpeg | Bin 245389 -> 0 bytes

analysis 4.R | 24 -----

analysis 5 output.jpeg | Bin 198610 -> 0 bytes

analysis 5.R | 32 -----

test.R | 21 ++++

15 files changed, 81 insertions(+), 217 deletions(-)

commit 9bb252c0b0daecc42ffa006bd9b05a6d79af099d

Author: nav334 <mn24act@herts.ac.uk>

Date: Sun Nov 24 19:05:46 2024 +0000

Add files via upload

final ppt A321.pptx | Bin 0 -> 810386 bytes

1 file changed, 0 insertions(+), 0 deletions(-)

commit cf7cb237089d4367f0dd1f496fe4c8dbd4097172

Author: manichandu007 <mk24afg@herts.ac.uk>

Date: Sun Nov 24 15:29:39 2024 +0000

Add files via upload

analysis 1 output.jpeg | Bin 0 -> 269226 bytes

analysis 1.R | 40 ++++++

analysis 2 output.jpeg | Bin 0 -> 77350 bytes

analysis 2.R | 49 ++++++

analysis 3 output.jpeg | Bin 0 -> 56221 bytes

analysis 3.R | 17 ++++++

analysis 4 output.jpeg | Bin 0 -> 245389 bytes

analysis 4.R | 24 ++++++

analysis 5 output.jpeg | Bin 0 -> 198610 bytes

analysis 5.R | 32 ++++++

10 files changed, 162 insertions(+)

commit b7c012eafb1039c26c7b6aaf8b5a4e37309af99c

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Sun Nov 24 00:36:34 2024 +0000

few fixes

Introduction | 5 -----

1 file changed, 5 deletions(-)

commit bdd1fe3149abff6554838d1f820cddac65998047

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Sun Nov 24 00:24:23 2024 +0000

box plot and other output

Analysis.R | 13 ++++++

Introduction | 5 +++++

Output/bell curve.png | Bin 0 -> 5488 bytes

Output/box plot.png | Bin 0 -> 50648 bytes

README.md | 2 +-

5 files changed, 19 insertions(+), 1 deletion(-)

commit 8a9df0d2f87d8b1cebaf87de21dd896ceb02c1bf

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Sun Nov 24 00:09:28 2024 +0000

visualization

Analysis.R | 7 +++++-

Output/Rplot.png | Bin 0 -> 9294 bytes

2 files changed, 6 insertions(+), 1 deletion(-)

commit fc3b46700347529e8fc9ff9f21a834ac80eb7777

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Thu Nov 21 20:54:22 2024 +0000

introduction push

Introduction => Introduction.txt | 0

1 file changed, 0 insertions(+), 0 deletions(-)

commit 4759afca12ff564bf006a1fbbd29a74fbb1c8f11

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Thu Nov 21 20:52:03 2024 +0000

minor fix on readme

README.md | 11 ++++++++--

1 file changed, 10 insertions(+), 1 deletion(-)

commit a000b8684f9d3c9df4c1182d8ac13c61ca9ea33a

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Thu Nov 21 20:49:38 2024 +0000

readme update

README.md | 1 +

1 file changed, 1 insertion(+)

commit 5f60bd40e0ddfe36ba5a971eff0882a0fe72bda9

Author: chukwuMonday <cm24abv@herts.ac.uk>

Date: Thu Nov 21 20:47:56 2024 +0000

Update README.md

README.md | 46 ++++++++-----

1 file changed, 15 insertions(+), 31 deletions(-)

commit c7a74d95559a36d1cf7c1b6ee8c92912f11e4fbd

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Thu Nov 21 20:41:56 2024 +0000

project push

Analysis.R

54

+++++

GROUP A321.pptx | Bin 0 -> 209311 bytes

Introduction | 5 +++++

README.md | 38 ++++++

4 files changed, 96 insertions(+), 1 deletion(-)

commit d69b40f08d7efd1f2ad04af5593e747140c439c5

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Thu Nov 21 18:25:24 2024 +0000

adding dataset

russian_demography.csv

2381

+++++

1 file changed, 2381 insertions(+)

commit 2ff3113e3aa3e8c3b8fc3c3fcf10a6263ec9753e

Author: Chukwu Monday <cm24abv@herts.ac.uk>

Date: Wed Nov 20 10:28:02 2024 +0000

testing first push

.gitignore | 4 +++++

TeamResearch.Rproj | 13 ++++++

2 files changed, 17 insertions(+)

commit 06db7fda4b276db543836775e4e01b01419ff277

Author: chukwuMonday <cm24abv@herts.ac.uk>

Date: Wed Nov 20 10:09:48 2024 +0000

Initial commit

README.md | 1 +

1 file changed, 1 insertion(+)