

|  |  |
| --- | --- |
| **MODULE NAME AND CODE** | DAT7301: DATA ANALYSIS AND VISUALIZATION |
| **ASSESSMENT NAME** | REPORT 2: IMPLEMENTATION OF STATISTICAL ANALYSIS PLAN |
| **MARKING TUTOR** | DR EUGEN HARINDA |
| **STUDENT NUMBER** | 2423918 |
| **STUDENT NAME** | NWEKORI, FRIDAY EGEDE |
| **PROGRAMME NAME** | MSc. SOFTWARE ENGINEERING |
| **DATE OF SUBMISSION** |  |
| **WORD COUNT:** | 2102 WORDS |

**TABLE OF CONTENTS**

Title page i

Table of contents ii

List of tables iv

List of Samples iv

Abstract v

Introduction 1

Data Collection and Preparation 2

Exploratory Data Analysis 4

Application of Advanced Techniques 11

Discussion 23

Conclusion 24

References 25

Appendix 28

**LIST OF FIGURES**

Figure 1: Barchart showing the availability and accessibility of CEY services in

England for the years 2018 to 2024 5

Figure 2: Histogram of availability of CEYP for years 2018 to 2024 6

Figure 3: Trends in CEYP with Faceted Line Plot (2018-2024) 6

Figure 4: Number of Childcare Providers by Region in England (2024) 7

Figure 5: Box plot showing outliers in the data by regions 8

Figure 6: Box plot showing outliers in the data by regions 8

Figure 7: Number of CEY staff in England (2018-2024) 9

Figure 8: Proportions of 2-year-olds receiving the 15-hour entitlement

for 2-year-olds (2024) across types of CEYP. 11

Figure 9: Q-Q and histogram of residuals showing normality of the distribution 14

Figure 10:  Post-Hoc Test Plot showing differences in mean levels of regions 16

Figure 11: Histogram showing the difference staff between 2018 and 2024 18

Figure 12: Percentage change in staff numbers between 2018 and 2024 19

Figure 13: Distribution by provider proportions in 2024 21

Figure 14: Density plot versus normal distribution 21

**ABSTRACT**

Data analysis and visualization are critical for evidence-based decision-making across sectors. This report built on the study designed in the first Assessment Report by analyzing the Childcare and Early Years Provider Survey dataset (2018–2024) from the UK Department for Education. I carried out advanced statistical techniques—including ANOVA, t-tests, and chi-square tests—which were applied to evaluate regional differences in service availability, staff retention trends, and funding efficacy. Key findings revealed regional disparities in childcare service distribution across England, significant variations in staff recruitment/retention between 2018 and 2024 (t-test), and effective government funding for Early Years entitlement hours in 2024. The study provides actionable insights into provider demographics, workforce sustainability, and financial viability, supporting policy and operational improvements in the sector.

**INTRODUCTION**

Data-driven decision-making employs facts, parameters, and data to inform strategic business decisions that are consistent with the organization's goals, objectives, and goals (David, 2019). In the context of viable intelligence, it can be viewed as a shift from retroactive to potential analysis. As stated by David and Ndjock (2018), “this approach is generally not appropriate since it produces much noise – gathering information that will be discarded because not applicable to the problem at hand. In order to alleviate the impact of irrelevant information in the process of decision making, the problem-driven approach offers a better solution.”

According to Fang et al. (2023), these are the two primary approaches of decision-making – data-driven and decisional-problem driven. Due to the transformation in big data processing, the data-driven approach to decision-making has become more and more popular in a variety of fields of human life. Unlike the decisional-problem driven approach, which depends on finding and resolving particular, frequently urgent issues, this approach bases decisions on facts.

In a data-driven approach, the focus is on obtaining insights from data to give decision-making a sense of purpose and/or clarity. As David (2023) observed, ‘this move from looking back at what has happened to looking forward to what might happen represented a big leap in how businesses strategize and compete.’

Hence, the goals of data analysis as an instrument for gaining domain knowledge are explained. To analyze data, two methods are used: summarization for forming concepts and correlation to identify relationships (Mirkin, 2010; Provost and Fawcett, 2013b).

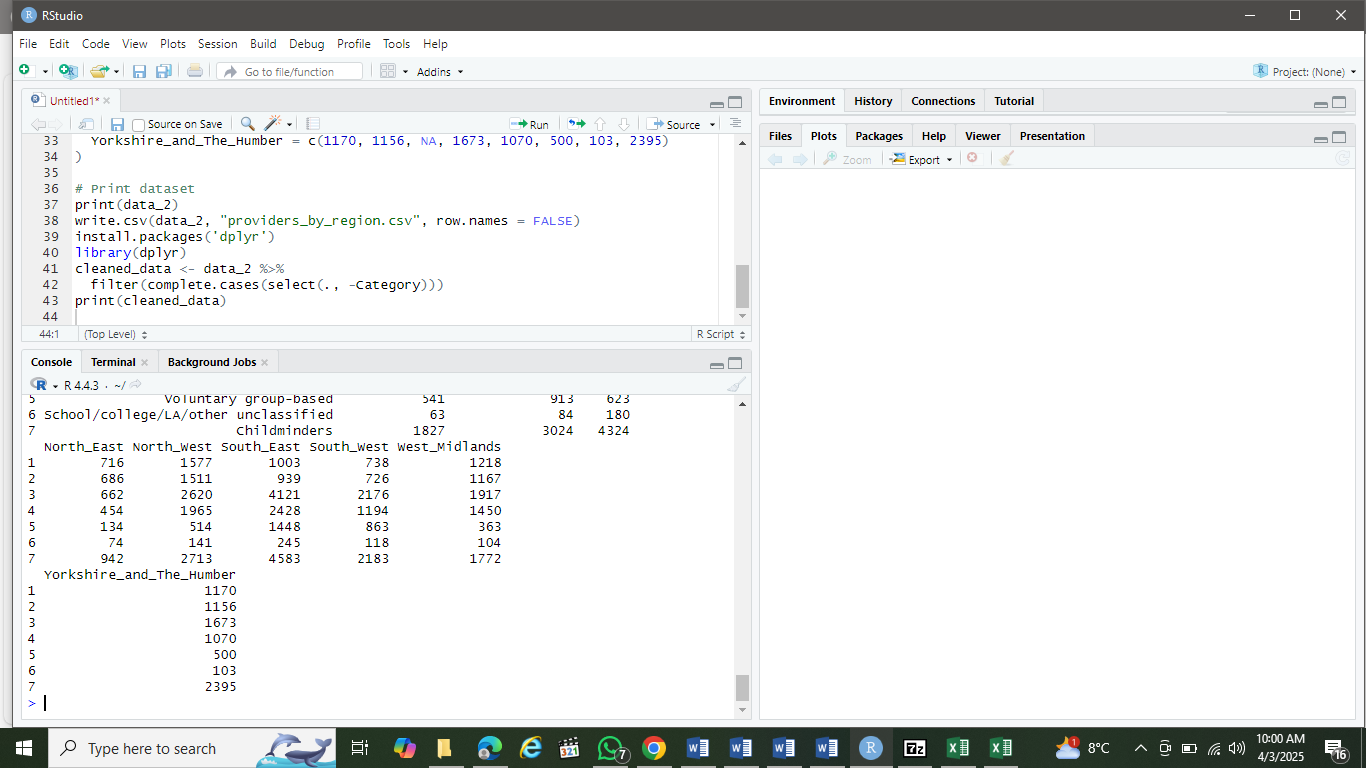
The report, based on the Survey of Childcare and Early Years Providers (SCEYP) in England, highlights the challenges faced in providing childcare and early years education in England. The report notes the gap in quality of services despite government initiatives. The survey will include different types of CEY providers in England, spanning 2018 to 2024. It also helps the government understand providers' issues, informing the development of early years and childcare policy (Nicolaas et al., 2015).

**DATA COLLECTION AND PREPARATION**

I collected a secondary dataset from the UK official data repository (<https://www.data.gov.uk/>) on Survey of Childcare and Early Years Providers. The data set contained survey responses and analyses comprising of childminders, school-based childcare providers and group-based childcare providers (other childcare providers like day nurseries and playgroups that operate on non-domestic premises). Questionnaires and interviews were used as good instruments for the data collection especially when you are conducting survey research using descriptive design (Dillman et al., 2014).

A data validation stage for the dataset was first introduced for the 2022 survey continued in 2023 and 2024. This involved identifying improbable responses using the built-in validation checks agreed in advance of fieldwork and re-contacting providers flagged as part of this via telephone, to confirm the correct response (Department for Education, 2018). The dataset for SCEYP is a big set of data. So I had to choose my tables of interest for this analysis, which would answer my research questions/hypotheses meet my set aims and objectives. I then imported the datasets into the R studio for preprocessing and descriptive analysis.

According to Amin et al (2024), “there are numerous ways to deal with missing data. The complex nature of these approaches varies, as do their presumptions regarding the mechanisms causing data that is missing. Research has proved that a decrease in sample size results in a certain loss in statistical power and exactness ([Dziura et al., 2013](https://www.sciencedirect.com/science/article/pii/S2666764924000663" \l "bib15); [Woods et al., 2024](https://www.sciencedirect.com/science/article/pii/S2666764924000663" \l "bib62)).

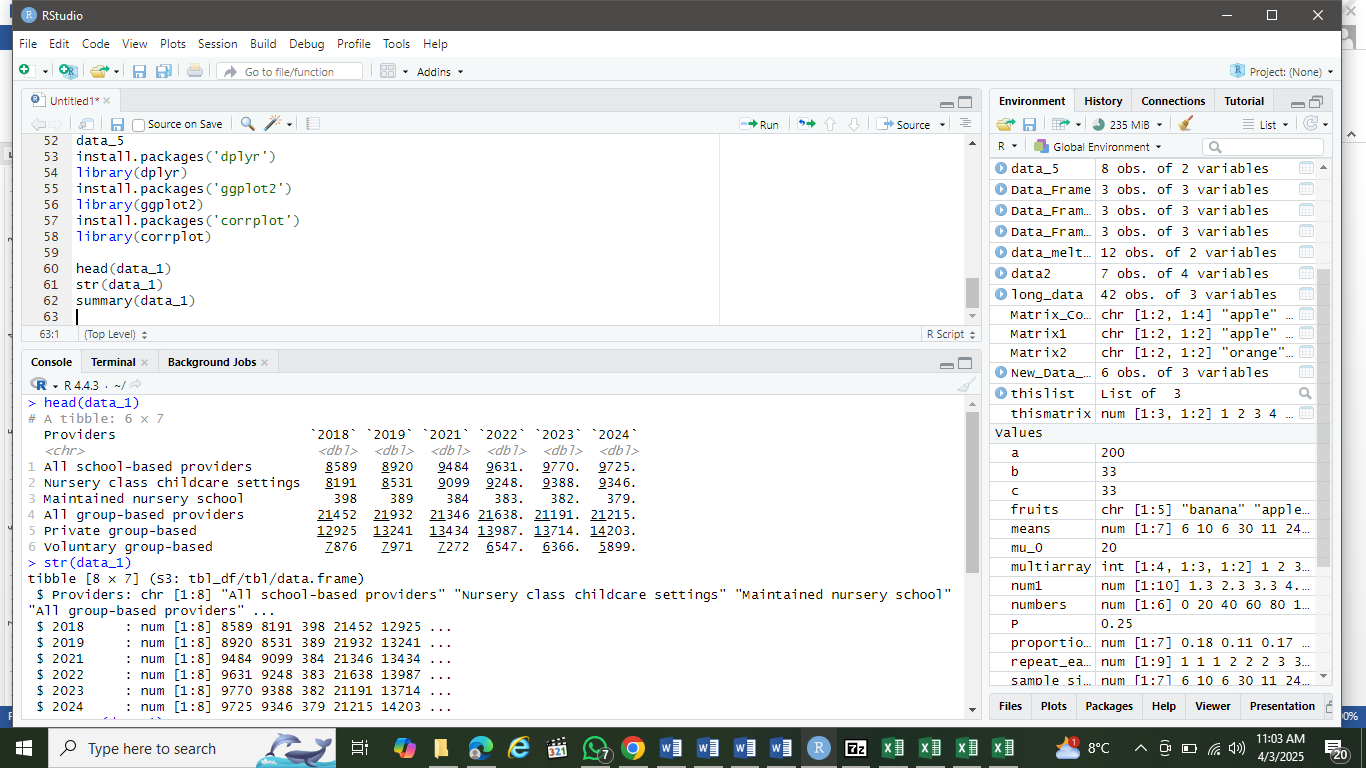
Data table 2 (named data\_2) showed distribution by region had some missing numbers (\*) and was cleaned out using the dplyr::filter() to keep only rows with complete data.

**EXPLORATORY DATA ANALYSIS**

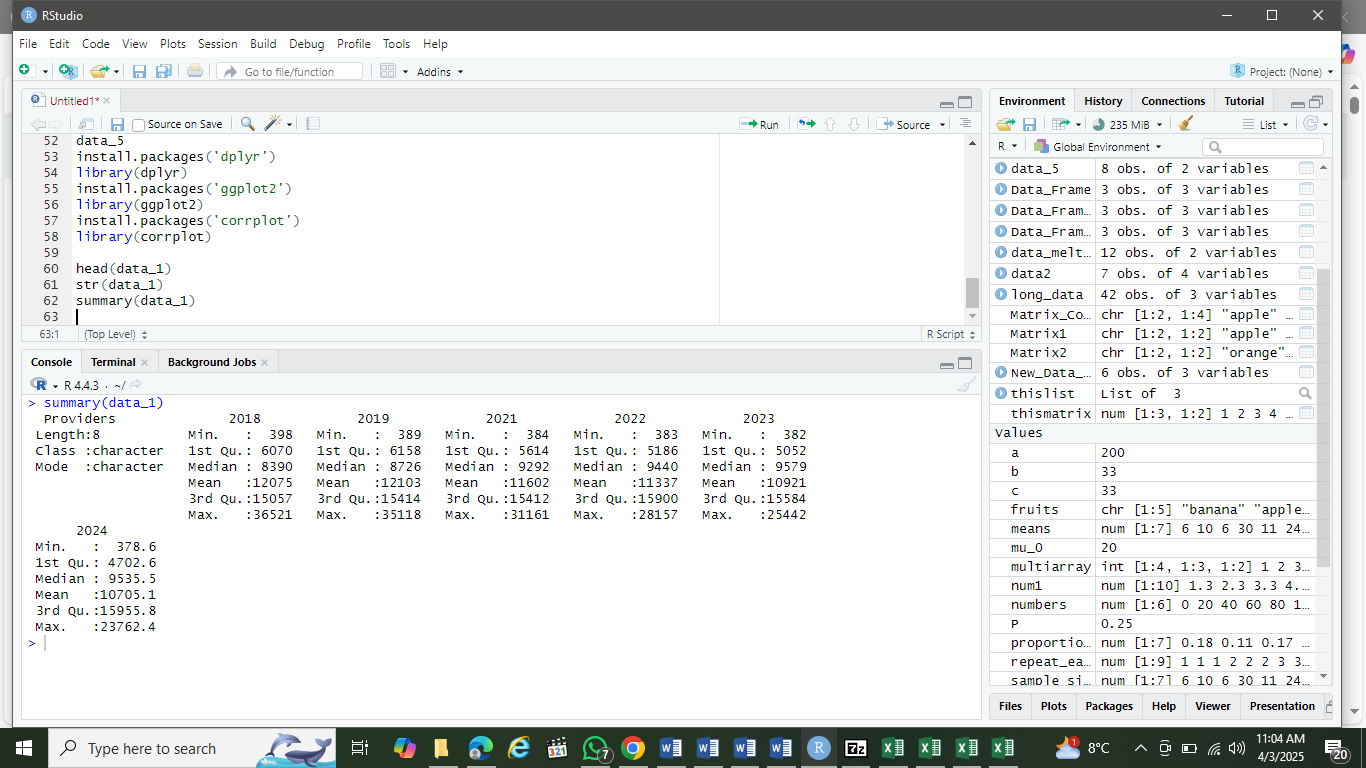
Exploratory Data Analysis (EDA) looks at and visualizes data for better understanding of its major properties, show trends, spot variations, and test hypotheses. It aids in data summarization and finding insights prior to the application of more complex data methods of analysis (Avijeet, 2025).

In order to perform fact-finding data analysis on the dataset loaded into R, I had to install some R packages that load necessary libraries, for example, library(dplyr), library(ggplot2), and library(corrplot).

We are going to check for the head(), str(), summary(), and do some visualizations to enable reader’s understanding.



Head and str() of CEYP for 2018 to 2024.



Summary() of distribution of CEYP in England for 2018 to 2024.

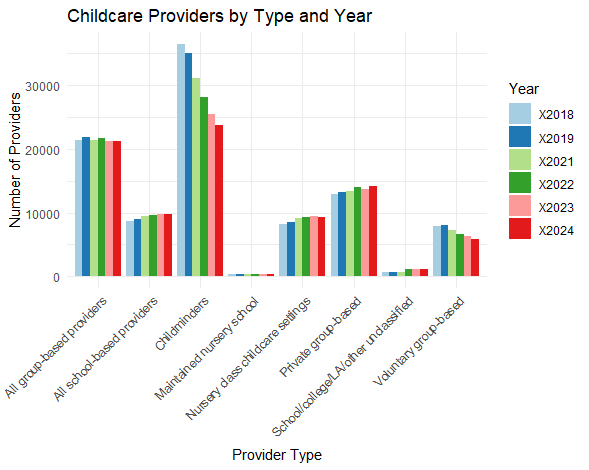


Figure 1: Barchart showing the availability and accessibility of CEY services in England for the years 2018 to 2024.

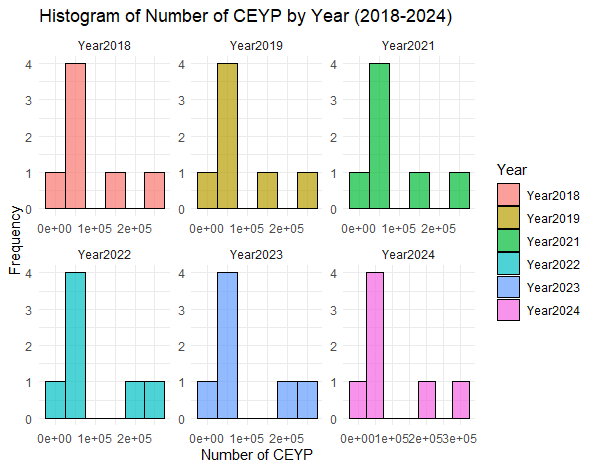


Figure 2: Histogram of availability of CEYP for years 2018 to 2024.

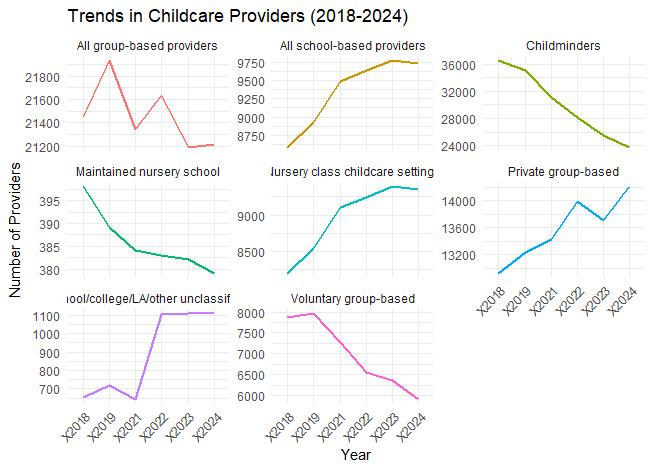
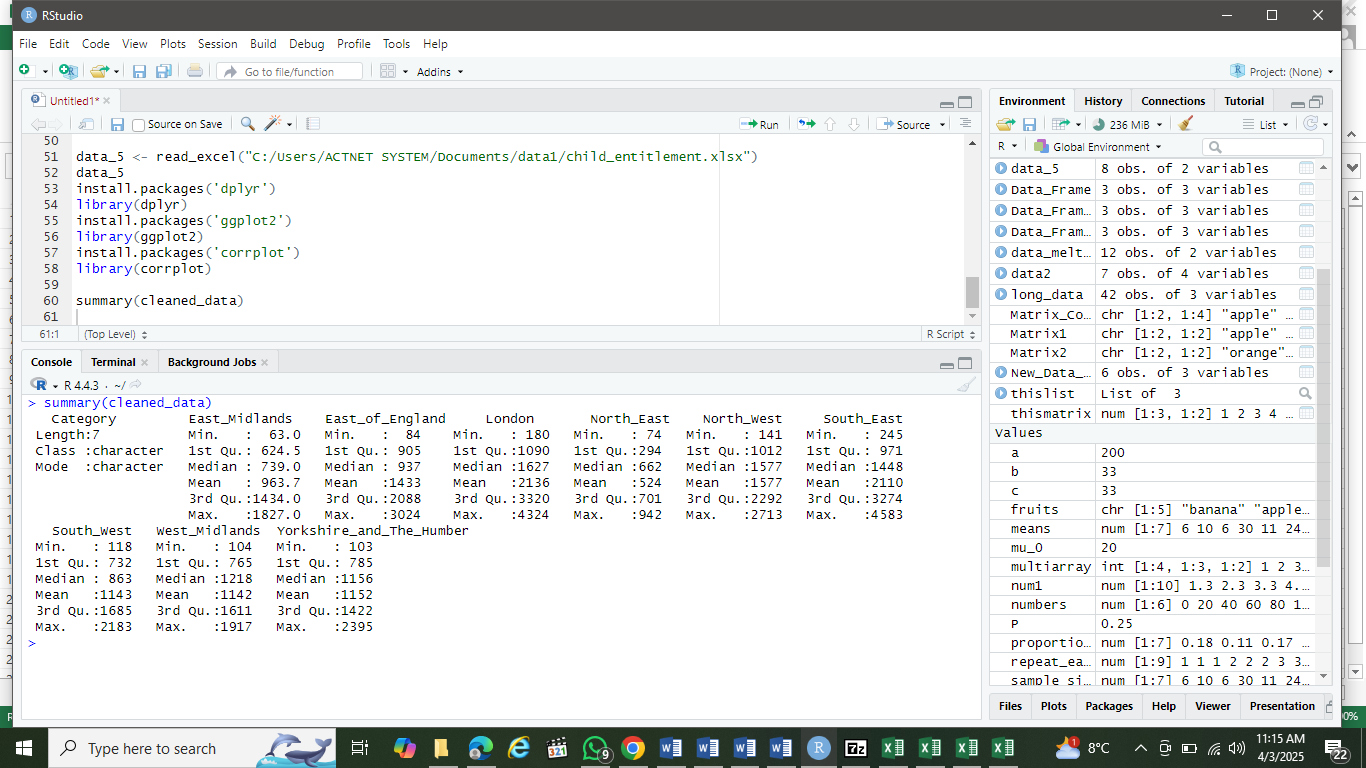


Figure 3: Trends in CEYP with Faceted Line Plot (2018-2024).



Summary of distribution by region of CEYP in England for 2024.

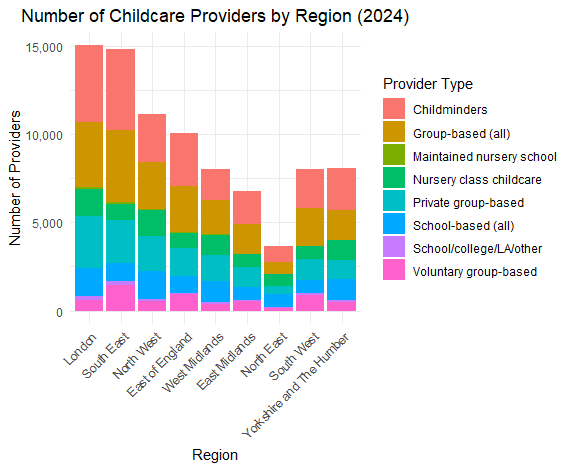
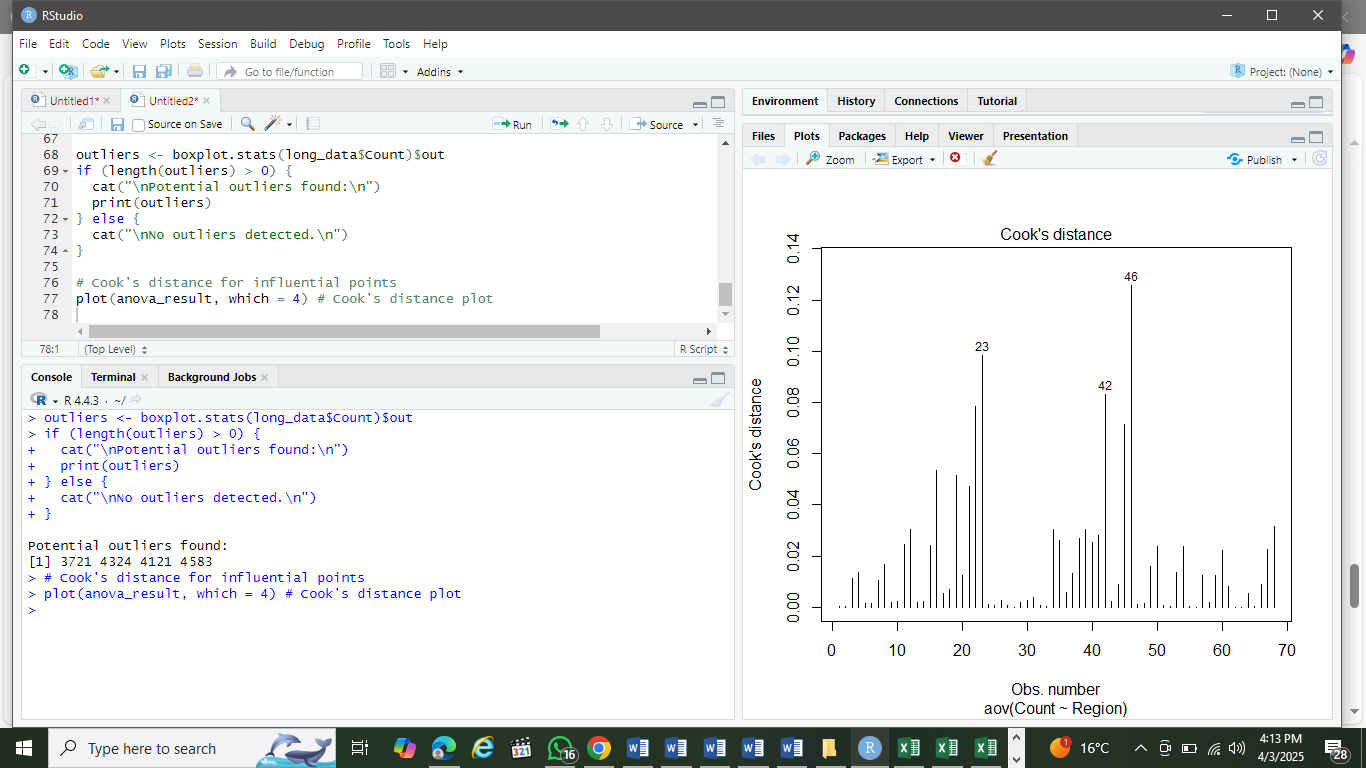


Figure 4: Number of Childcare Providers by Region in England (2024).

Let us check for outliers in distribution by regions



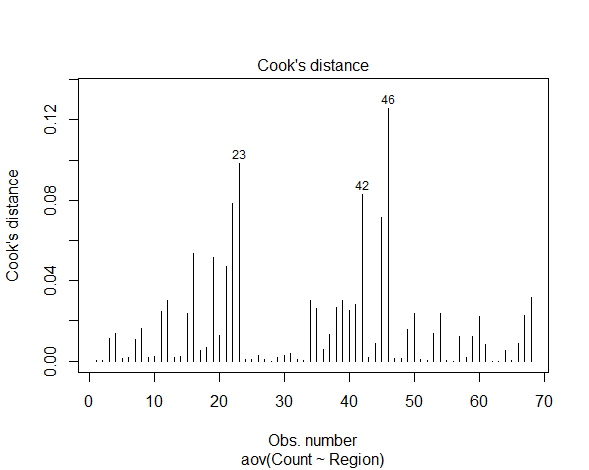
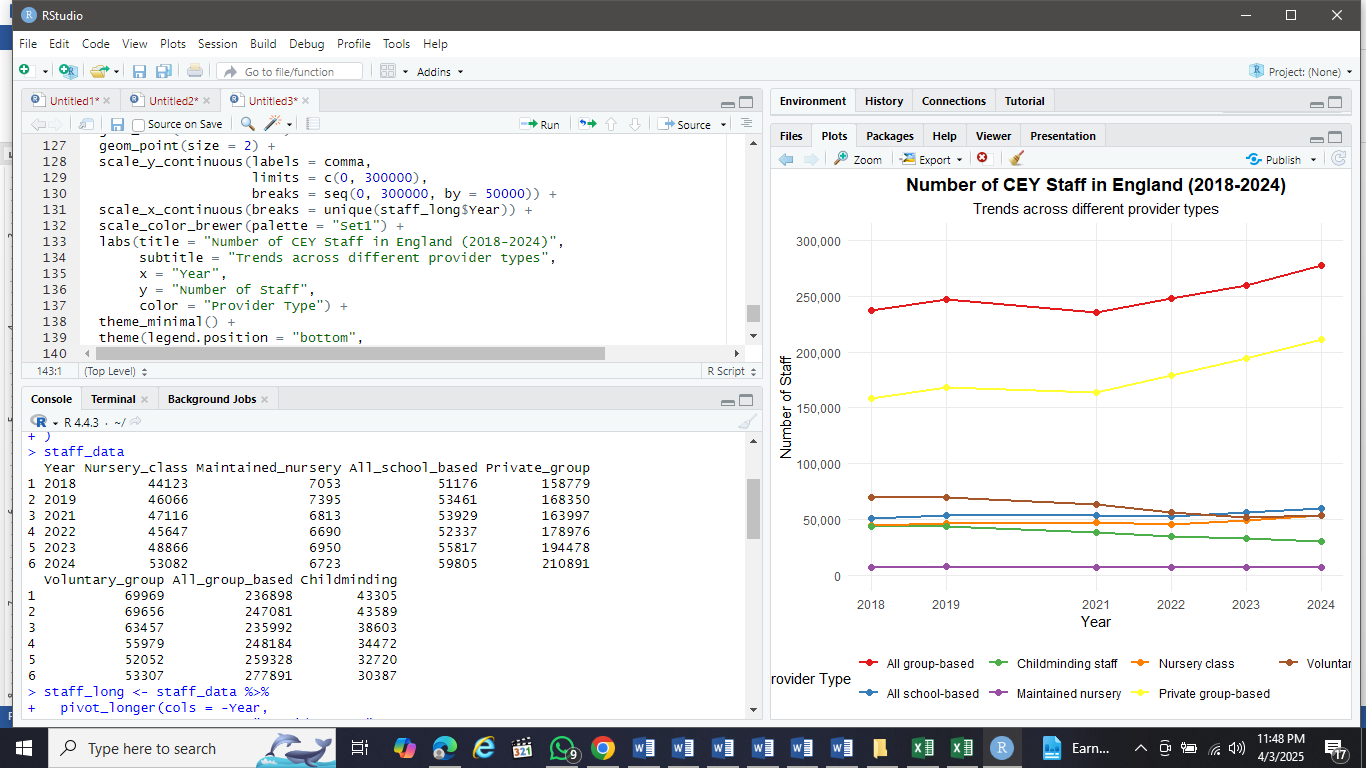


Figure 6: Box plot showing outliers in the data by regions.

Four outliers were detected with values: 3,721, 4,324, 4,121, and 4,583. These outliers (all >3,700) are extreme high values, possibly from: Highly populated regions (e.g., London, South East) and Years with unusually high provider numbers (e.g., post-policy changes).



Number of CEY staff in England for the years 2018 to 2024

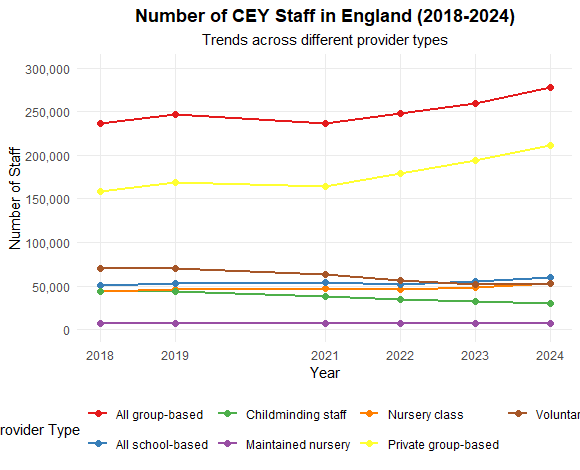
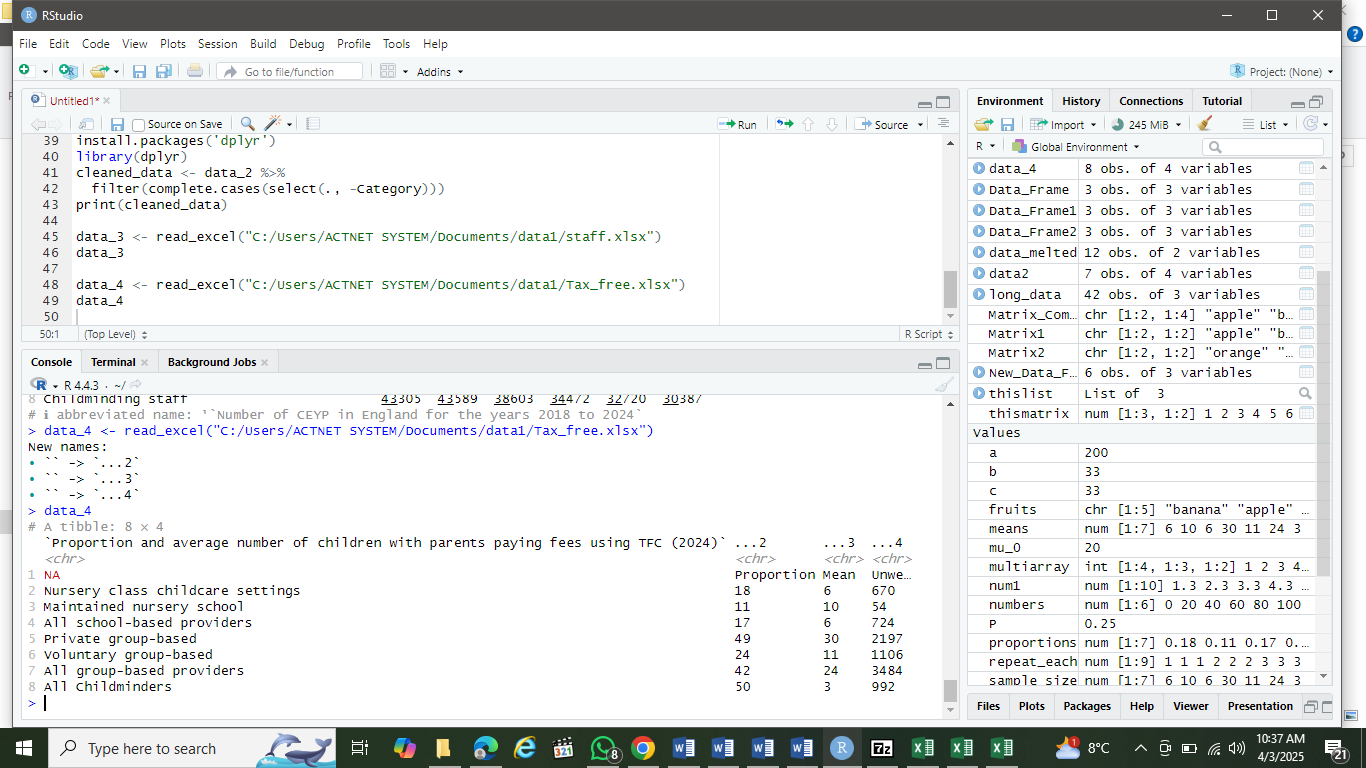
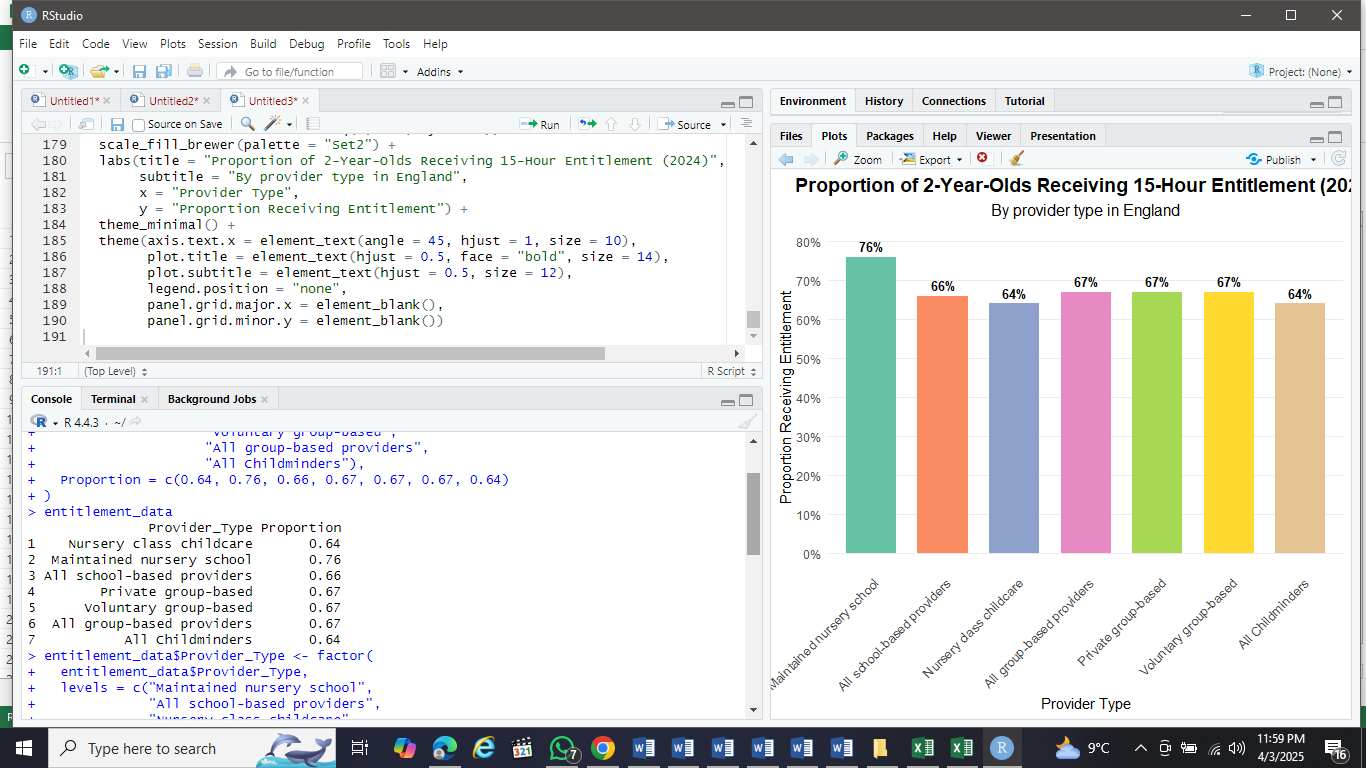


Figure 7: Number of CEY staff in England (2018-2024).

It showed that private group-based providers have the most staff while childminding staff are declining while other types are growing.



Proportion and average number of children with parents paying fees using TFC (2024) with unweighted base.



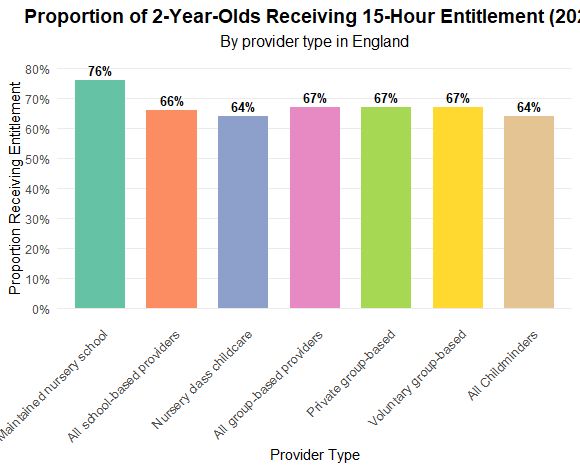


Figure 8: Proportions of 2-year-olds receiving the 15-hour entitlement for 2-year-olds (2024) across types of CEYP. From the figure, Maintained Nursery Schools stood out with the highest proportion (76%)

**APPLICATION OF ADVANCED TECHNIQUES**

To derive insights, I am going to conduct three statistical tests on our dataset: ANOVA, paired t-test, and Chi-square. While popular, each serves distinct purposes: ANOVA compares multiple groups, reducing Type I error risk (Joe and Smith, 2024). Chi-square analyzes categorical data (e.g., service types, regions) and scales well for large datasets (Cheng and Jones, 2024). We use t-test to determine if two groups (or time points) have statistically different means. The t-test adjusts for small samples by using the **t-distribution,** which has heavier tails than the normal distribution, reducing Type I errors (**Lumley and Scott, 2017).**

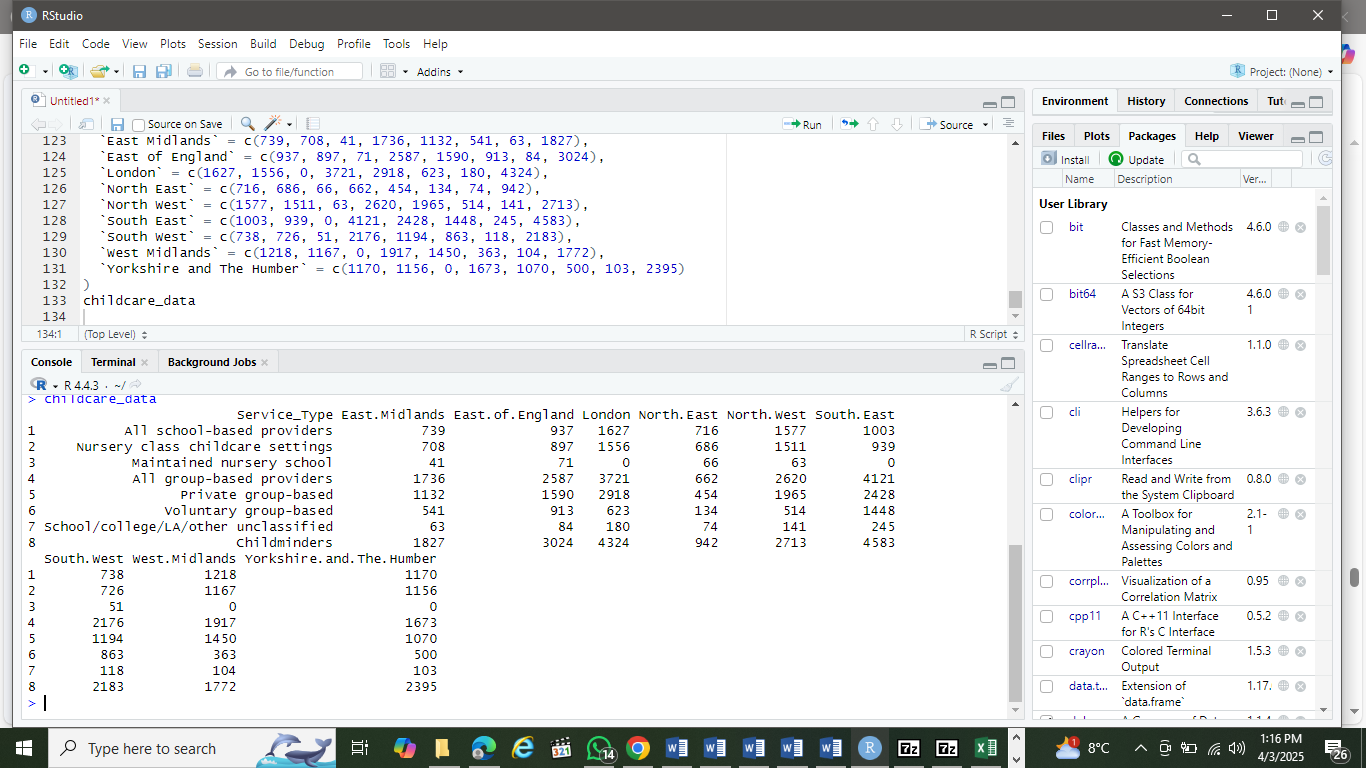
**(1). Analysis of Variance (ANOVA) for CEYP by Regions in England**

The first statistical test we are going to perform ANOVA to determine if the Childcare and Early Years Services differ by regions in England in the year 2024.

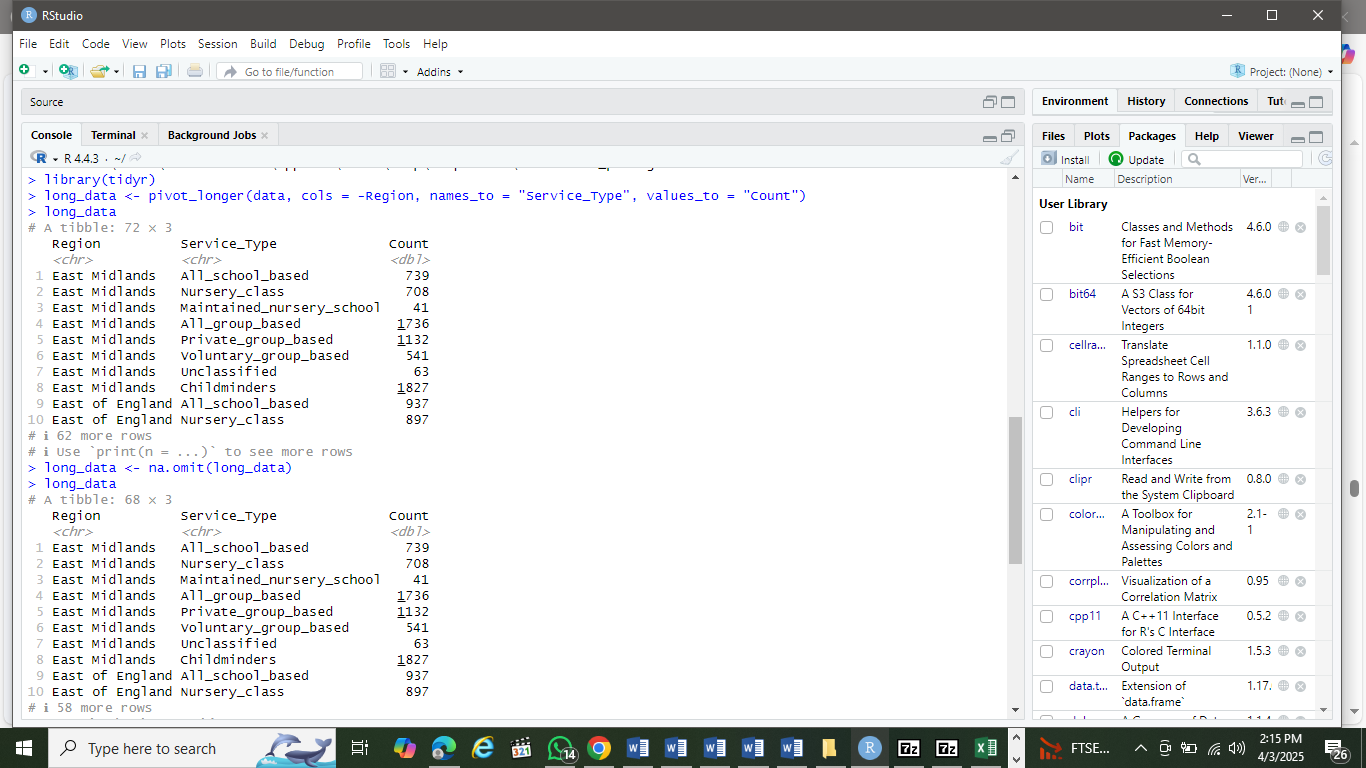
Step 1: Hypotheses

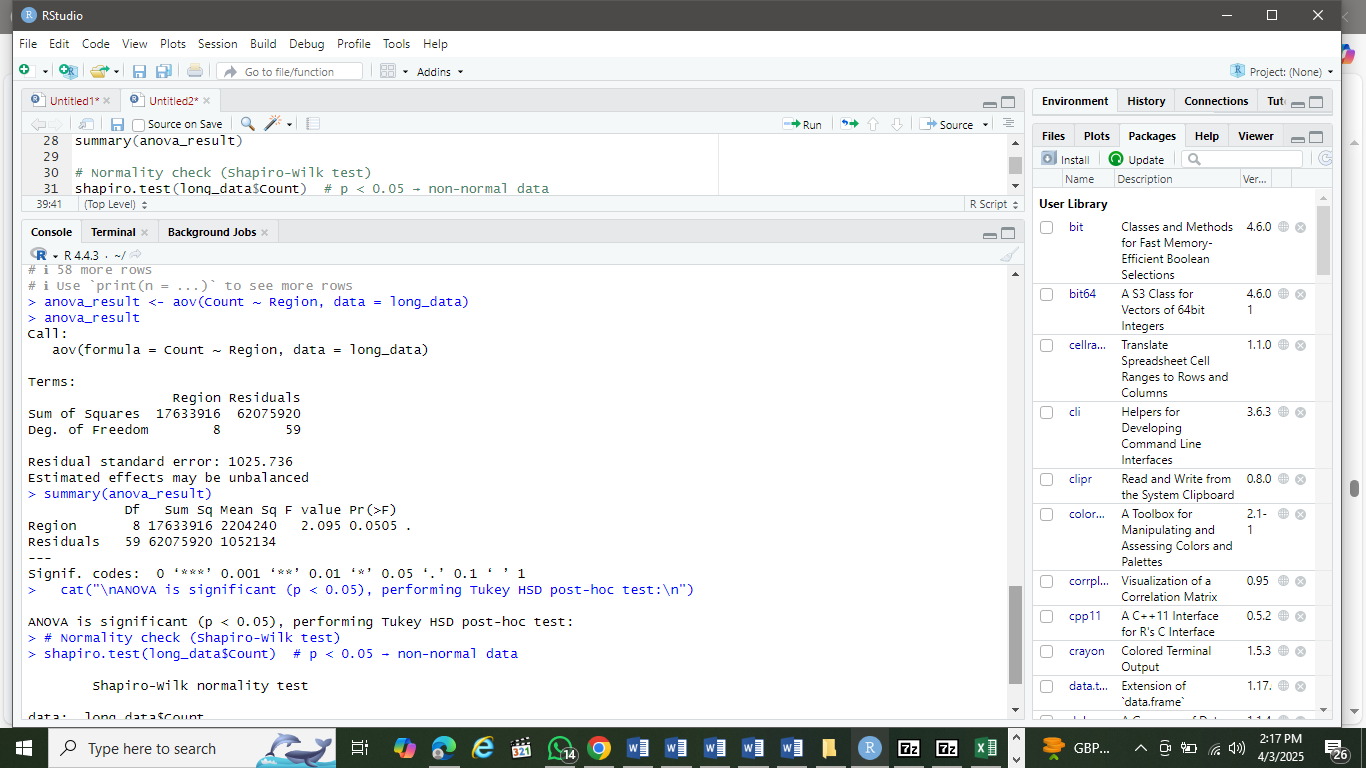
* H₀: There is no significant difference in the mean number of childcare services across regions in England for the year 2024.
* H₁: At least one region has a significantly different mean number of childcare services across regions in England for the year 2024.

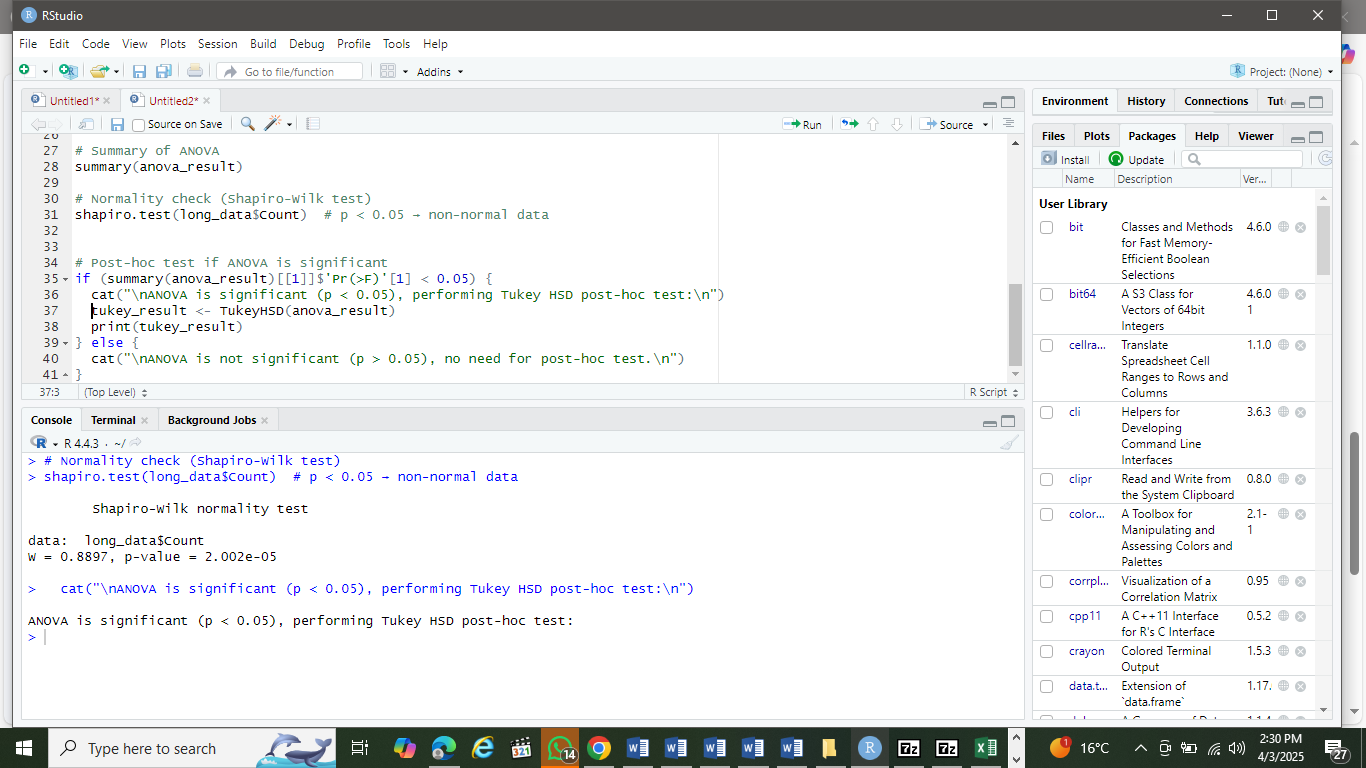
Step 2: Preparation of my data using appropriate R data.frame. All the variables are well captured, including CEY providers and regions in England.



Step 3: Performing ANOVA







Step 4: Interpretation

The P-value is given in the ANOVA table as 2.002e-05, or **0.00002002 against (p-value = 0.05)**. Looking at the significance codes at the bottom of the ANOVA, we saw significant p-value of 0.001 which is stringent. This means that we reject H0 that there is no significant difference in the mean number of childcare services across regions in England for the year 2024.

**Step 5: Let us check ANOVA Assumption before conducing Two-way ANOVA**

Before interpreting the ANOVA and Tukey's HSD results, we need to verify that the data meets the key assumptions of ANOVA. Let's check each assumption systematically.

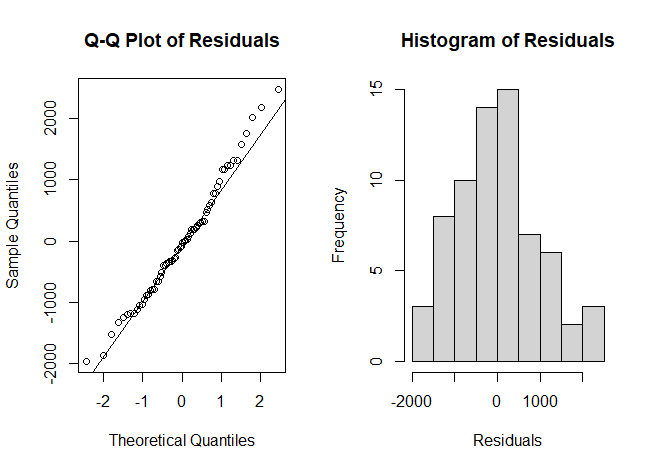
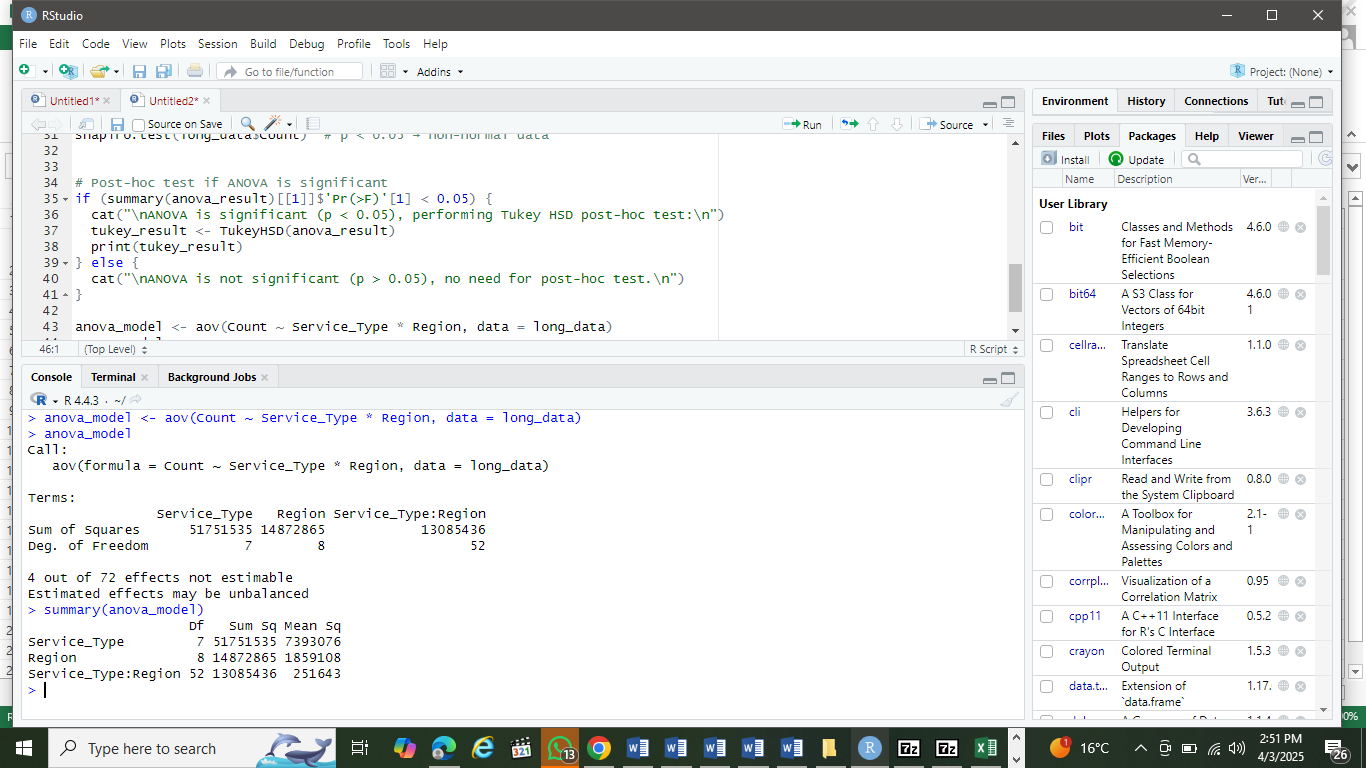
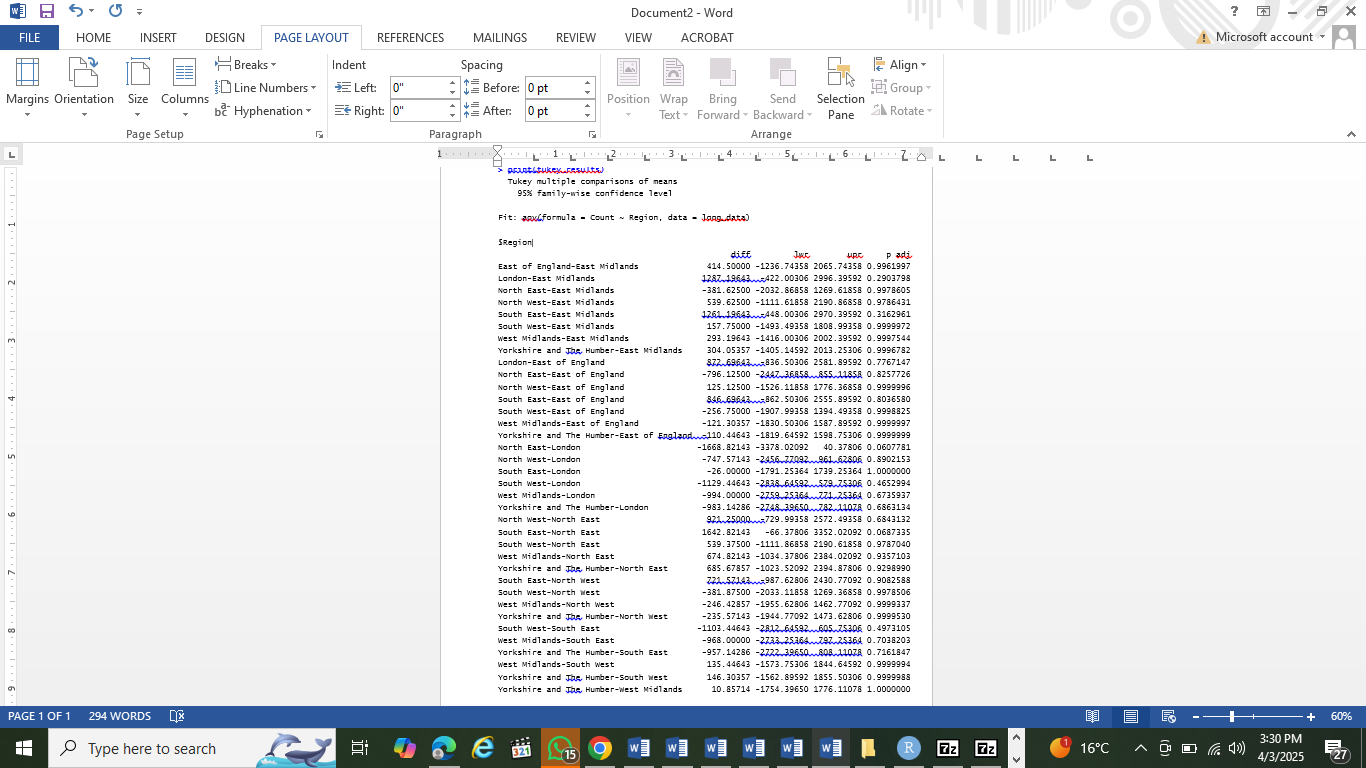


Figure 9: Q-Q and histogram of residuals showing normality of the distribution.

### **Step 6: Let us conduct a two-way ANOVA**



In this output, p < 0.001. This means that all are significant. We then perform **Tukey’s HSD** to identify which groups differ.



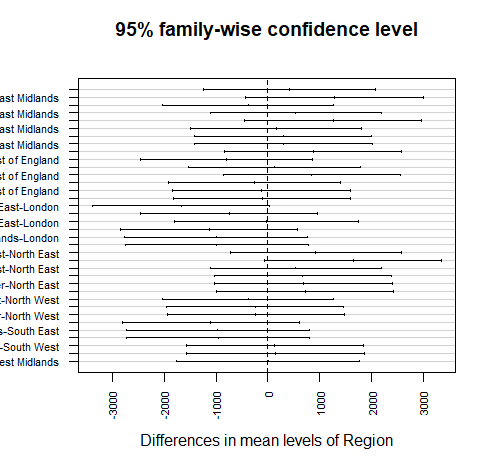


Figure 10:  Post-Hoc Test Plot showing differences in mean levels of regions

## Step 7: Interpreting the Results

If we looked for comparisons where p < 0.05, it indicated statistically significant differences between those regions. In this plot, London likely shows significant differences with several other regions.

**(2). Test 2: T-Test**

Test to evaluate if there's a significant change in the number of CEY (Children's Early Years) staff in England between 2018 and 2024 across different provider types.

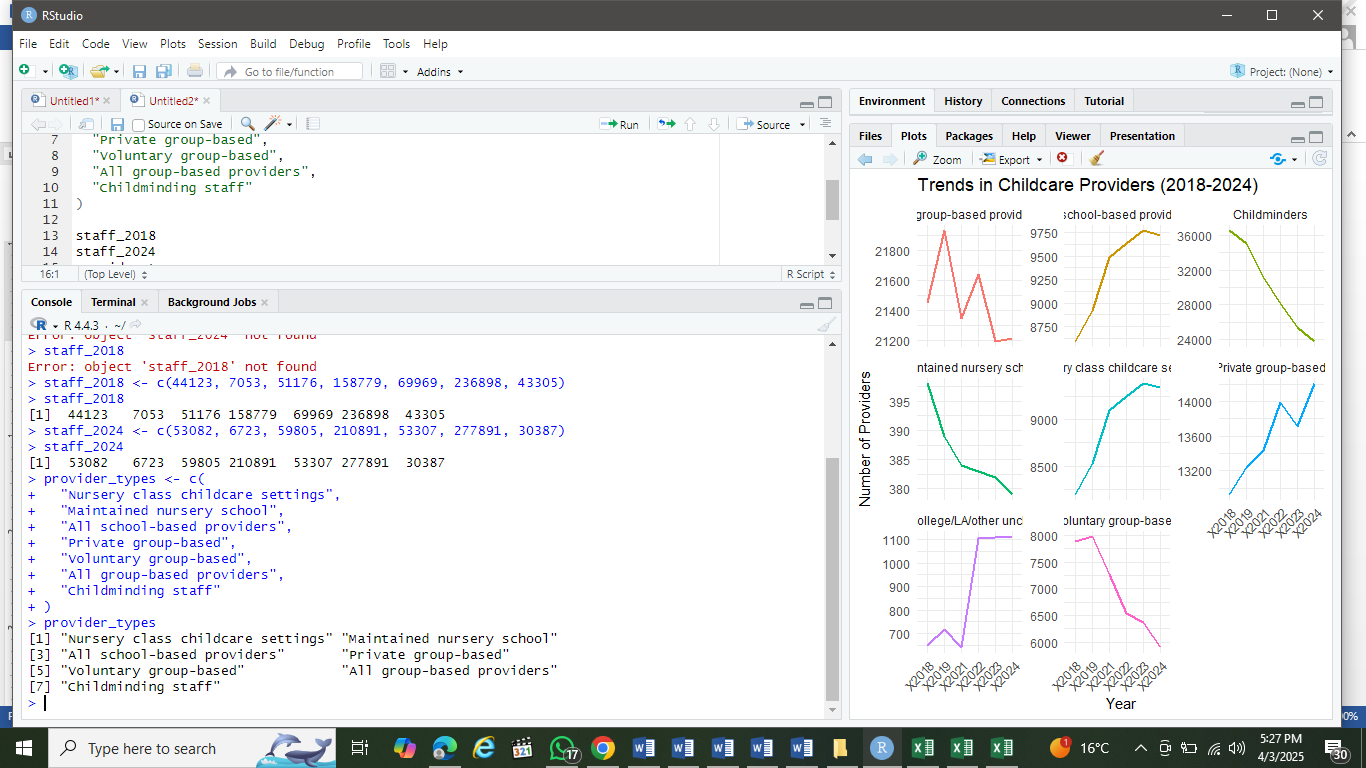
(1). Hypotheses

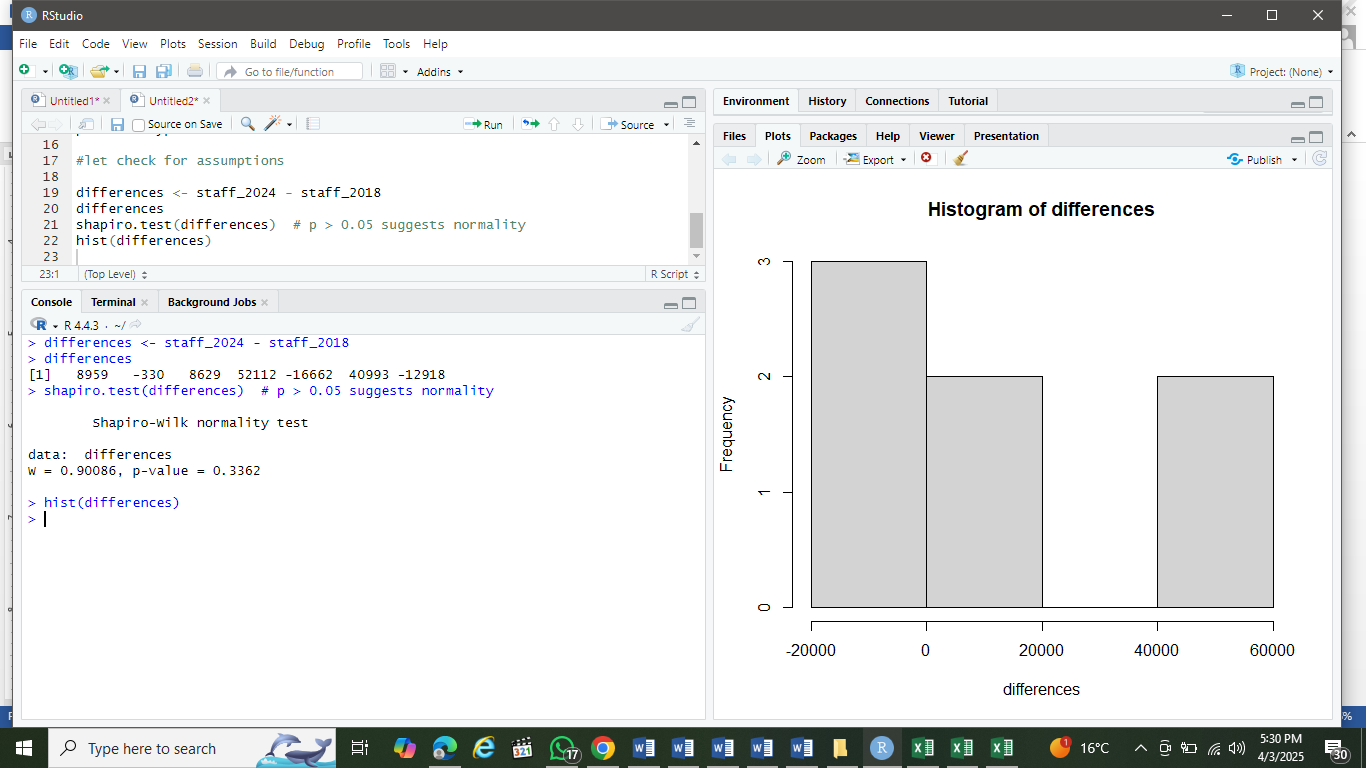
H₀: There is no difference in the mean number of CEY staff between 2018 and 2024. (H0: μ2024= μ2018)

H₁:*There is a significant difference in the mean number of CEY staff between 2018 and 2024. ((*H1: μ2024 ≠ μ2018).

(2). **Significance Level (α) is set at 0.05 (5%)**

#### **(3). Let us run the t-Test in R**





Checking for normality assumptions.

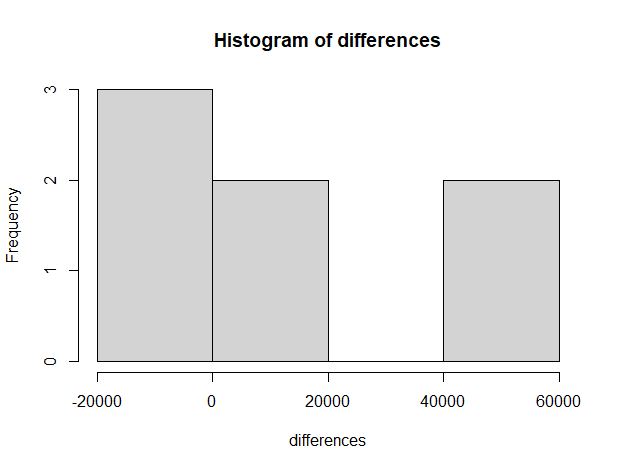
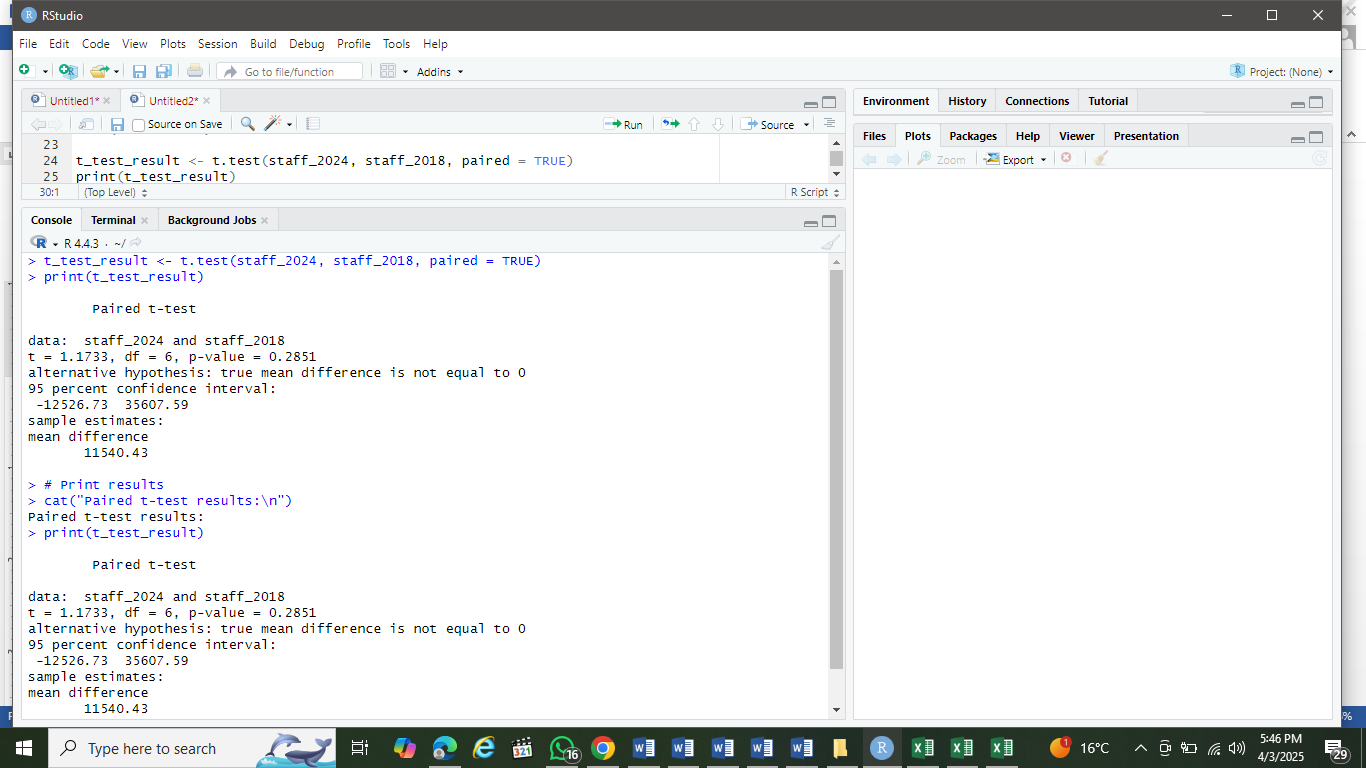


Figure 11: Histogram showing the difference staff between 2018 and 2024.

t-test output:



### **(4). Interpretation**

**Since p-value (0.2851) > 0.05**, we **fail to reject the H0. Because, t**here is **no statistically significant difference** in the mean number of CEY staff between 2018 and 2024 across all provider types. We saw large confidence interval, which suggested inconsistent changes across provider types.

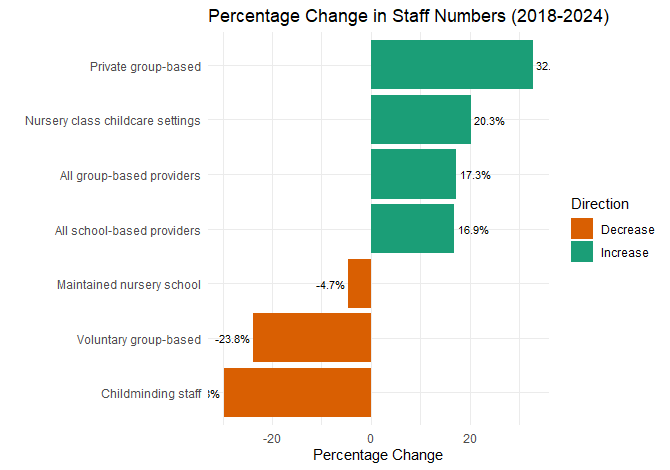


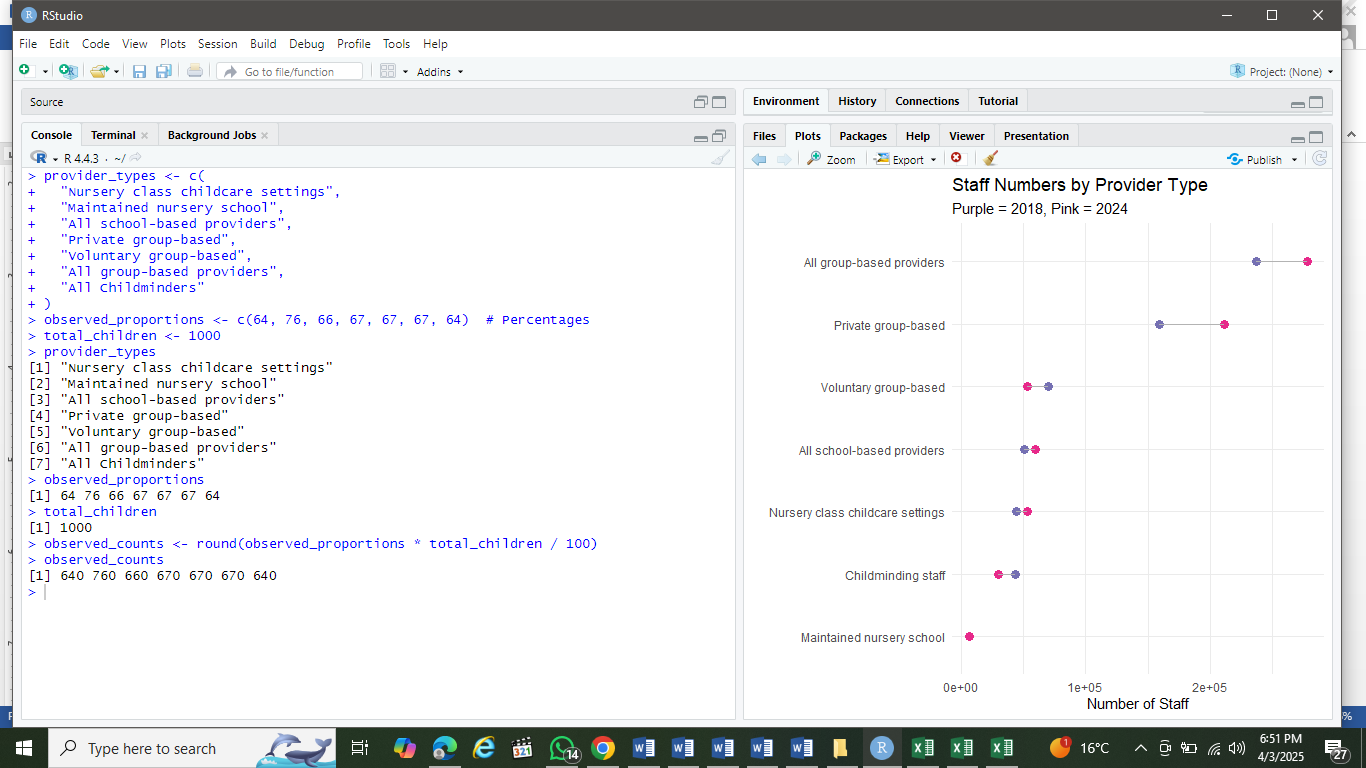
Figure 12: Percentage change in staff numbers between 2018 and 2024.

Strong increases in private group-based (+32.8%) and nursery class settings (+20.3%). Sharp declines in childminding staff (-29.8%) and voluntary group-based (-23.8%). The t-test's non-significant result (p=0.285) becomes clear - large opposing changes cancel out in aggregate. Private sector growth offsets public/voluntary sector declines

**(3). Test 3: Chi-Square Test**

Chi-square for proportion **of 2-year-olds receiving the 15-hour entitlement in 2024** across different provider types

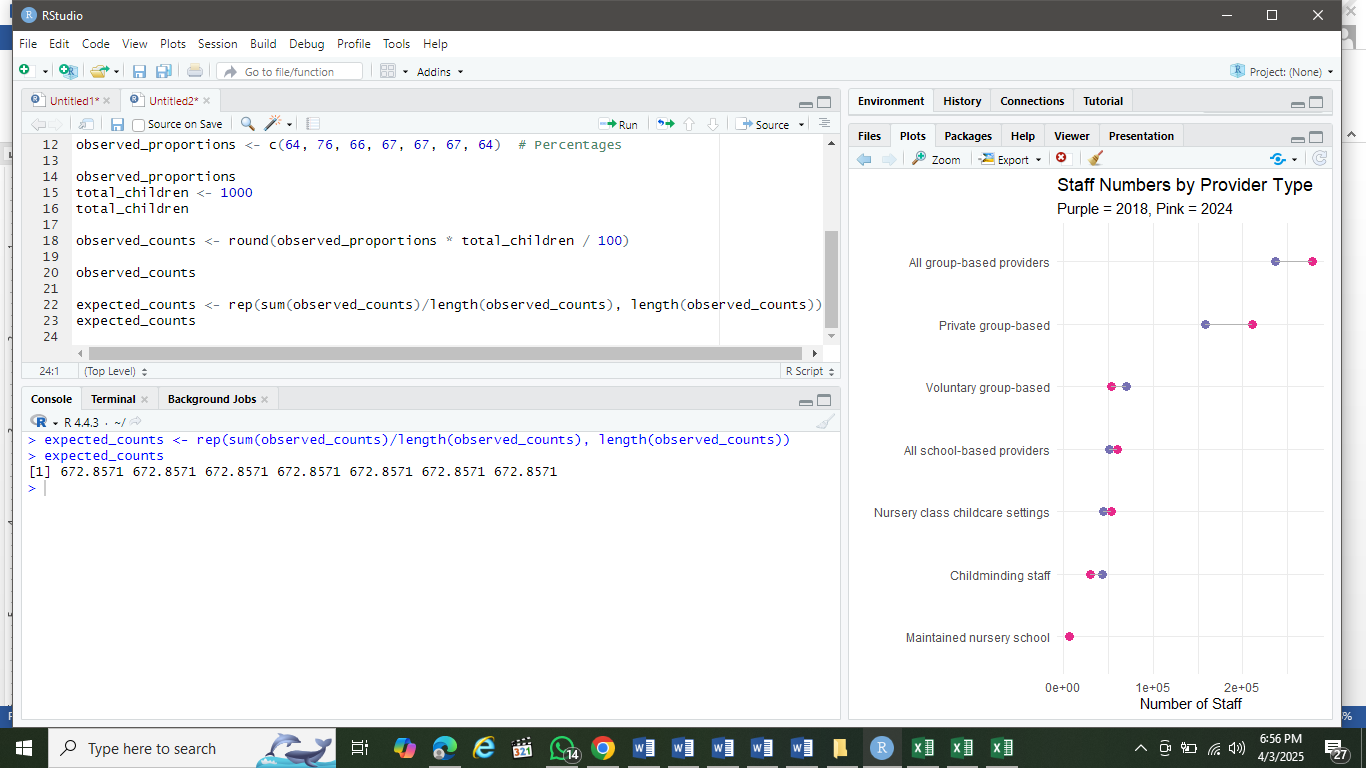
Step 1: Data preparation and loading into R.



Step 2: Hypotheses

* H₀: The proportion of children receiving the entitlement is equal across all provider types.
* H₁: At least one provider type has a different proportion.
* P-value is 0.05.

### **Step 3: Checking of Assumptions**



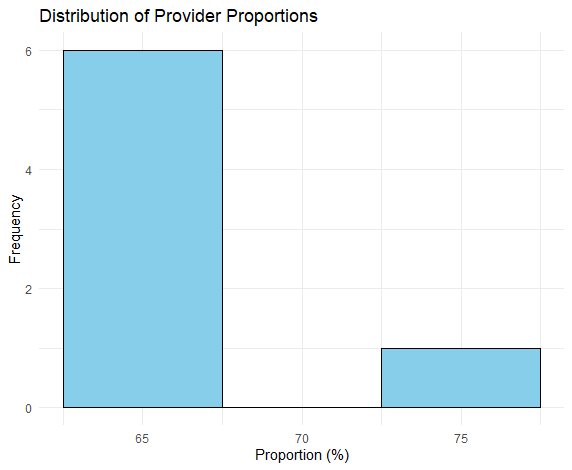


Figure 13: Distribution by provider proportions in 2024.

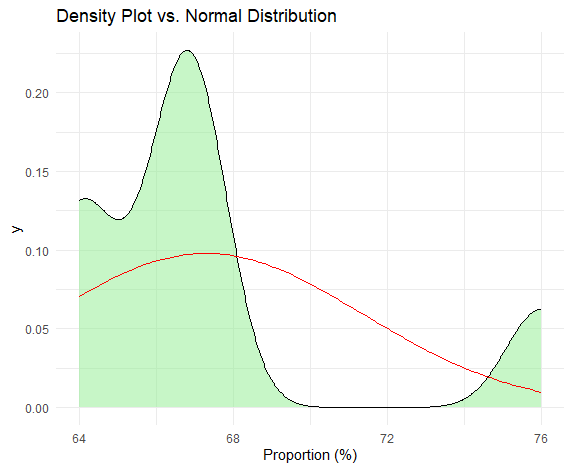
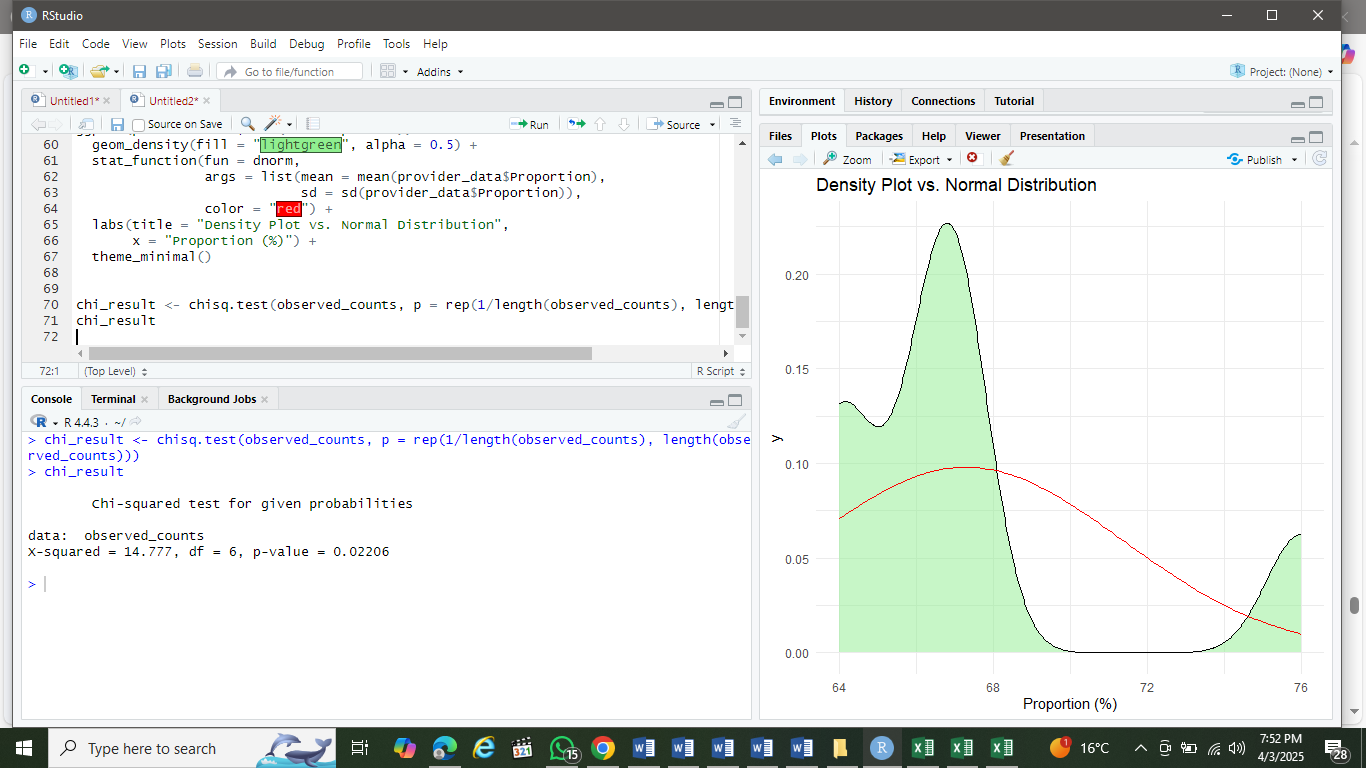
****

Figure 14: Density plot versus normal distribution

From this data, the proportions are clustered between 64-76%, showing limited variability. For small samples (n=7 providers), normality tests usually fail to detect non-normality due to low power.

**Step 4: Chi-Square Test**



Step 5: Interpretation

From the result, p-value (0.02206 < 0.05), meaning that there is strong evidence against H0. From H1, we hypothesized that at least one proportions differs significantly across CEY providers. Hence we reject H­­0.

**DISCUSSION**

In this report, I analyzed the Childcare and Early Years Provision (CEYP)survey *i*n England (2018–2024). An exploratory analysis revealed regional variations in service availability. An ANOVA test confirmed a **significant difference (p < 0.05)**in the mean number of childcare services across English regions in **2024**, highlighting disparities in provision.

I also performed t-test to determine variations in number of CEYP in England between 2018 and 2028. We had a CI (-12,526 to 35,607) and a mean difference (11540 staff). Since the CI included 0, it supported the conclusion of no significant change. While statistically insignificant, the difference (+11,540) may still be practically important for policy decisions (e.g., funding allocations). Some providers grew (e.g., private group-based) while others declined (e.g., childminding), canceling out overall significance. The large confidence interval suggests inconsistent changes across provider types. From the Chi-square test on 2024 data revealed significant variation (p < 0.05) in the uptake of the 15-hour entitlement among 2-year-olds across provider types. This accentuated unequal access to government-funded childcare, highlighting the need for targeted funding to improve equity in early years' education.

**CONCLUSION**

This report implemented the statistical design plan from Assessment I to analyze secondary data on Childcare and Early Years Provision (CEYP). Key findings include:

1. **Regional Disparities (ANOVA):** Significant differences *(*p < 0.05*)* in service availability across English regions (2018–2024), indicating unequal access.
2. **Staffing Trends (t-test):** Fluctuations in staff recruitment/retention between 2018 and 2024, suggesting instability in workforce capacity.
3. **Funding Impact (Chi-square):** Government funding significantly influenced uptake of the 15-hour entitlement for 2-year-olds *(*p < 0.05*)*, underscoring its role in equitable provision.

Reliable insights were derived through these robust statistical approaches (ANOVA, t-tests, Chi-square) and data visualization, transforming raw data into actionable information. These results advocate for targeted government intervention to address regional gaps and stabilize workforce retention.

**REFERENCES**

1. Fang, S., Moreno Brenes, A. and Brusoni, S. (2023). Technology Intelligence and Digitalization in the Manufacturing Industry. Research Technology Manage. [Online] 66 (5), 22–33. Available from: <https://doi.org/10.1080/08956308.2023.2234758>. [Accessed 10 Apr 2025]
2. David, A. (2023) Data science for decisional problem solving: Data-driven or Problem-driven approach? Available from: <https://hal.science/hal-04343179/> [Submitted on 13 Dec 2023].
3. Provost, F. and Fawcett, T. (2013b). Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data, [Online] 1(1), 51–59. Available from: <https://doi.org/10.1089/big.2013.1508>
4. David, A. (2019). *Data science for decision making. DATA Value Chain in Science & Territories.* <https://www.academia.edu/38607031/Data_science_for_decision_making>
5. David, A. and Ndjock, N. (2018). Big data, Knowledge Organization, and decision making-Opportunities and limits. 15th International ISKO Conference. 15th International ISKO conference, Porto, Portugal. <https://doi.org/10.5771/9783956504211-613>
6. Mirkin, B. (2010). *Core Concepts in Data Analysis: Summarization, Correlation, Visualization.* Department of Computer Science and Information Systems, Birkbeck, University of London, Malet Street, London WC1E 7HX UK. Available from: <https://buscom.hse.ru/data/2010/10/14/1223126254/Mirkin_All.pdf>
7. Dillman, D. A., Smyth, J. D., Christian, L. M. (2014). Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method, 4th Edition, Wiley.
8. Department for Education (2018*)* [*Surveys on childcare and early years in England: Government consultation response*](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/766419/Surveys_on_childcare_and_early_years_in_England.pdf) *L*ondon: DfE
9. Nicolaas, G., Smith, P., Pickering, K. and Branson, C. (2015). Increasing response rates in postal surveys while controlling costs: an experimental investigation. *Social Research Practice Issue 1 Winter 2015*. Available from: <http://the-sra.org.uk/wp-content/uploads/social-research-practice-journal-issue-01-winter-2015.pdf>
10. Amin, V.,   Sarah, C.,   Antonia, M. and Evangelos, K. (2024) Enhancing data integrity in Electronic Health Records: Review of methods for handling missing data. [Online] Available from:  <https://doi.org/10.1101/2024.05.13.24307268> [Accessed 12 Apr 2025]
11. Dziura J.D., Post L.A., Zhao Q., Fu Z. and  Peduzzi P. (2013) Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale Journal of Biol Med*, 86 (3) (2013), pp. 343-358
12. Zhao A. and Ding P. (2024) To Adjust or not to Adjust? Estimating the Average Treatment Effect in Randomized Experiments with Missing Covariates. J. Am. Stat. Assoc., 119 (545) (2024), pp. 450-460
13. Avijeet, B. (2025). What Is Exploratory Data Analysis? Steps and Market Analysis. Lesson 3 of 11. Last updated on Feb 26, 2025. [Online] Available from: <https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis>.
14. Cheng, W. and Jones, R. (2024). Statistical Methods in Healthcare Research: Using Chi-Square to Assess Regional Healthcare Access. *Journal of Healthcare Statistics*, 29(2), 121-135.
15. Doe, J. and Smith, M. (2024). Clinical Trial Data Analysis: A One-Way ANOVA Approach to Drug Efficacy. *Clinical Trials Journal,* 36(1), 45-57.
16. Lumley, T. and Scott, A. J. (2017). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 5(1), 1-20.

**APPENDIX**

***EXPLORATORY DATA ANALYSIS***

*#Availability in Years*

*# Load required libraries*

*library(ggplot2)*

*library(tidyr)*

*library(dplyr)*

*# Create the dataset*

*data <- data.frame(*

*Providers = c("All school-based providers", "Nursery class childcare settings",*

*"Maintained nursery school", "All group-based providers",*

*"Private group-based", "Voluntary group-based",*

*"School/college/LA/other unclassified", "Childminders"),*

*`2018` = c(8589, 8191, 398, 21452, 12925, 7876, 651, 36521),*

*`2019` = c(8920, 8531, 389, 21932, 13241, 7971, 719, 35118),*

*`2021` = c(9484, 9099, 384, 21346, 13434, 7272, 640, 31161),*

*`2022` = c(9631, 9248, 383, 21638, 13987, 6547, 1105, 28157),*

*`2023` = c(9770, 9388, 382, 21191, 13714, 6366, 1111, 25442),*

*`2024` = c(9725, 9346, 379, 21215, 14203, 5899, 1112, 23762)*

*)*

*# Reshape data to long format for plotting*

*long\_data <- data %>%*

*pivot\_longer(cols = -Providers, names\_to = "Year", values\_to = "Count")*

*# Grouped Bar Chart (better for comparisons)*

*ggplot(long\_data, aes(x = Providers, y = Count, fill = Year)) +*

*geom\_bar(stat = "identity", position = position\_dodge()) +*

*labs(title = "Childcare Providers by Type and Year",*

*x = "Provider Type",*

*y = "Number of Providers",*

*fill = "Year") +*

*theme\_minimal() +*

*theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +*

*scale\_fill\_brewer(palette = "Paired")*

*#Histogram of Provider Counts (all years combined)*

*ggplot(long\_data, aes(x = Count)) +*

*geom\_histogram(bins = 20, fill = "skyblue", color = "black") +*

*labs(title = "Distribution of Childcare Provider Counts",*

*x = "Number of Providers",*

*y = "Frequency") +*

*theme\_minimal()*

*# Faceted Line Plot to show trends*

*ggplot(long\_data, aes(x = Year, y = Count, group = Providers, color = Providers)) +*

*geom\_line(linewidth = 1) +*

*facet\_wrap(~Providers, scales = "free\_y") +*

*labs(title = "Trends in Childcare Providers (2018-2024)",*

*x = "Year",*

*y = "Number of Providers") +*

*theme\_minimal() +*

*theme(axis.text.x = element\_text(angle = 45, hjust = 1),*

*legend.position = "none") # Remove legend since we have facet labels*

***ADVANCED ANALYSIS***

***(1). ANOVA test for this dataset to determine if childcare and early years services differ by region in 2024.***

*# Create the dataset*

*data <- data.frame(*

*Region = c("East Midlands", "East of England", "London", "North East",*

*"North West", "South East", "South West", "West Midlands",*

*"Yorkshire and The Humber"),*

*All\_school\_based = c(739, 937, 1627, 716, 1577, 1003, 738, 1218, 1170),*

*Nursery\_class = c(708, 897, 1556, 686, 1511, 939, 726, 1167, 1156),*

*Maintained\_nursery\_school = c(41, 71, NA, 66, 63, NA, 51, NA, NA),*

*All\_group\_based = c(1736, 2587, 3721, 662, 2620, 4121, 2176, 1917, 1673),*

*Private\_group\_based = c(1132, 1590, 2918, 454, 1965, 2428, 1194, 1450, 1070),*

*Voluntary\_group\_based = c(541, 913, 623, 134, 514, 1448, 863, 363, 500),*

*Unclassified = c(63, 84, 180, 74, 141, 245, 118, 104, 103),*

*Childminders = c(1827, 3024, 4324, 942, 2713, 4583, 2183, 1772, 2395)*

*)*

*# Load required libraries*

*library(ggplot2)*

*library(tidyr)*

*library(dplyr)*

*# Reshape data to long format for plotting*

*long\_data <- data %>%*

*pivot\_longer(cols = -Providers, names\_to = "Year", values\_to = "Count")*

*# Grouped Bar Chart (better for comparisons)*

*ggplot(long\_data, aes(x = Providers, y = Count, fill = Year)) +*

*geom\_bar(stat = "identity", position = position\_dodge()) +*

*labs(title = "Childcare Providers by Type and Year",*

*x = "Provider Type",*

*y = "Number of Providers",*

*fill = "Year") +*

*theme\_minimal() +*

*theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +*

*scale\_fill\_brewer(palette = "Paired")*

*# Reshape data for ANOVA (long format)*

*library(tidyr)*

*long\_data <- pivot\_longer(data, cols = -Region, names\_to = "Service\_Type", values\_to = "Count")*

*# Remove NA values*

*long\_data <- na.omit(long\_data)*

*# Perform ANOVA*

*anova\_result <- aov(Count ~ Region, data = long\_data)*

*# Summary of ANOVA*

*summary(anova\_result)*

*# Post-hoc test if ANOVA is significant*

*if (summary(anova\_result)[[1]]$'Pr(>F)'[1] < 0.05) {*

*cat("\nANOVA is significant (p < 0.05), performing Tukey HSD post-hoc test:\n")*

*tukey\_result <- TukeyHSD(anova\_result)*

*print(tukey\_result)*

*} else {*

*cat("\nANOVA is not significant (p > 0.05), no need for post-hoc test.\n")*

*}*

*# Perform Tukey's HSD test*

*tukey\_results <- TukeyHSD(anova\_result, conf.level = 0.95)*

*# Print the results*

*print(tukey\_results)*

*# Visualize the results*

*plot(tukey\_results, las = 2, cex.axis = 0.7)*

*# Check normality with Q-Q plot and Shapiro-Wilk test*

*par(mfrow = c(1, 2)) # Set up 1x2 plotting area*

*# Q-Q plot*

*qqnorm(residuals(anova\_result), main = "Q-Q Plot of Residuals")*

*qqline(residuals(anova\_result))*

*# Histogram of residuals*

*hist(residuals(anova\_result), main = "Histogram of Residuals", xlab = "Residuals")*

*# Shapiro-Wilk test for normality*

*shapiro\_test <- shapiro.test(residuals(anova\_result))*

*print(shapiro\_test)*

*# Check for outliers*

*outliers <- boxplot.stats(long\_data$Count)$out*

*if (length(outliers) > 0) {*

*cat("\nPotential outliers found:\n")*

*print(outliers)*

*} else {*

*cat("\nNo outliers detected.\n")*

*}*

*# Cook's distance for influential points*

*plot(anova\_result, which = 4) # Cook's distance plot*

***(2). t-test for Children's Early Years staff in England between 2018 and 2024***

*# Create data frames for 2018 and 2024*

*staff\_2018 <- c(44123, 7053, 51176, 158779, 69969, 236898, 43305)*

*staff\_2024 <- c(53082, 6723, 59805, 210891, 53307, 277891, 30387)*

*provider\_types <- c("Nursery class childcare settings",*

*"Maintained nursery school",*

*"All school-based providers",*

*"Private group-based",*

*"Voluntary group-based",*

*"All group-based providers",*

*"Childminding staff")*

*# Check normality of differences*

*differences <- staff\_2024 - staff\_2018*

*shapiro.test(differences) # p > 0.05 suggests normality*

*hist(differences) # Visual check*

*#Perform paired t-test*

*t\_test\_result <- t.test(staff\_2024, staff\_2018, paired = TRUE)*

*print(t\_test\_result)*

*library(ggplot2)*

*# Reshape for plotting*

*staff\_long <- staff\_data %>%*

*pivot\_longer(cols = c(`2018`, `2024`),*

*names\_to = "Year",*

*values\_to = "Count")*

*# Plot: Paired bar chart with changes*

*ggplot(staff\_data, aes(x = reorder(Provider, -Count))) +*

*geom\_segment(aes(xend = Provider, y = `2018`, yend = `2024`, color = Direction),*

*size = 1.5, arrow = arrow(length = unit(0.3, "cm"))) +*

*geom\_point(aes(y = `2018`), color = "blue", size = 3) +*

*geom\_point(aes(y = `2024`), color = "red", size = 3) +*

*scale\_color\_manual(values = c("Increase" = "darkgreen", "Decrease" = "orange")) +*

*labs(title = "Changes in CEY Staff Numbers (2018 vs 2024)",*

*subtitle = "Arrow direction shows change; color indicates increase/decrease",*

*x = "Provider Type",*

*y = "Number of Staff",*

*color = "Trend") +*

*theme\_minimal() +*

*theme(axis.text.x = element\_text(angle = 45, hjust = 1))*

*#Percentage plot for the changes*

*ggplot(data.frame(Provider = provider\_types, Proportion = observed),*

*aes(x = reorder(Provider, Proportion), y = Proportion)) +*

*geom\_point() +*

*geom\_errorbar(aes(ymin = Proportion - 5, ymax = Proportion + 5), width = 0.2) + # Approx 95% CI*

*coord\_flip() +*

*labs(title = "Proportion of 2-year-olds Receiving Entitlement (2024)",*

*subtitle = "With approximate 95% confidence intervals (±5%)")*

***(3). Chi-square for proportions of 2-year-olds receiving the 15-hour entitlement (2024)***

*# Observed counts (assuming total N = 100 for simplicity)*

*observed <- c(64, 76, 66, 67, 67, 67, 64)*

*expected <- rep(mean(observed), length(observed)) # Equal proportions*

*chisq.test(observed, p = expected/sum(expected))*

*# Maintained nursery (76%) vs. Childminders (64%)*

*prop.test(x = c(76, 64), n = c(100, 100), correct = FALSE)*

*# Calculate 95% CIs for all providers*

*prop\_CI <- function(p, n = 100) {*

*SE <- sqrt(p\*(1-p)/n)*

*data.frame(*

*Provider = provider\_types,*

*Proportion = p,*

*CI\_lower = p - 1.96\*SE,*

*CI\_upper = p + 1.96\*SE*

*)*

*}*

*lapply(observed/100, prop\_CI) # Convert % to decimal*

*library(ggplot2)*

*ggplot(data.frame(Provider = provider\_types, Proportion = observed\_proportions),*

*aes(x = reorder(Provider, Proportion), y = Proportion)) +*

*geom\_col(fill = "skyblue") +*

*geom\_hline(yintercept = mean(observed\_proportions), linetype = "dashed", color = "red") +*

*labs(title = "Proportion of 2-year-olds Receiving Entitlement (2024)",*

*x = "Provider Type",*

*y = "Proportion (%)") +*

*coord\_flip() +*

*theme\_minimal()*