Ayush Pradhan

Retail Sales Analysis: Store layout and signage impact on sales:

Exploratory Data Analysis:

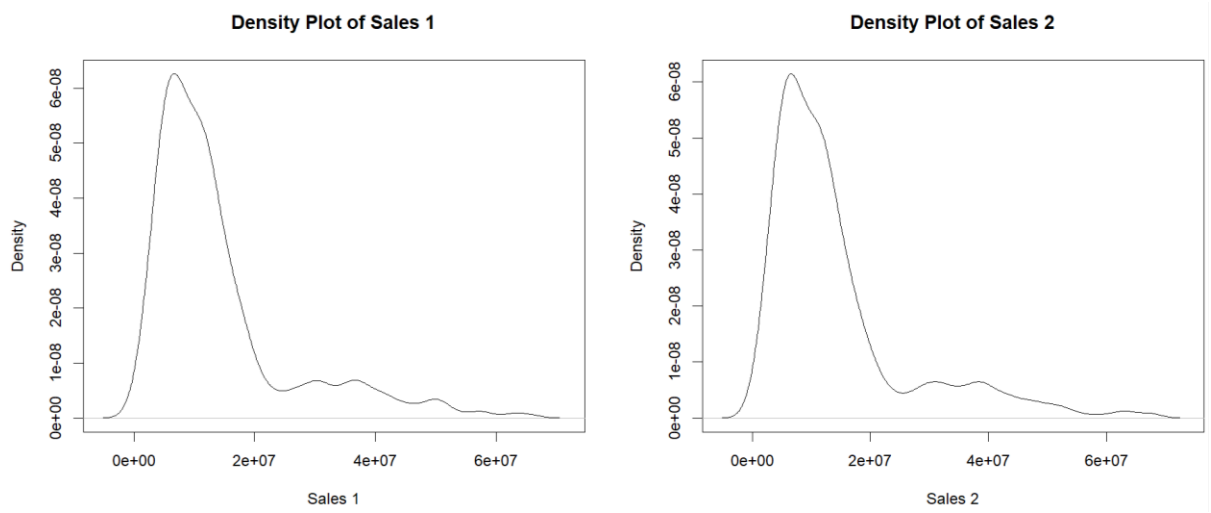


Figure 1: Density plots of Sales 1 and Sales 2, apparent right-skewness

The density plots of sales 1 and sales 2 in Figure 1 provide insights into the distribution of sales before and after the layout and signage change. Both sales 1 and sales 2 appear to be right-skewed, with a few stores having significantly higher sales than the majority. This suggests that a transformation of the sales variables might be necessary for the analysis.

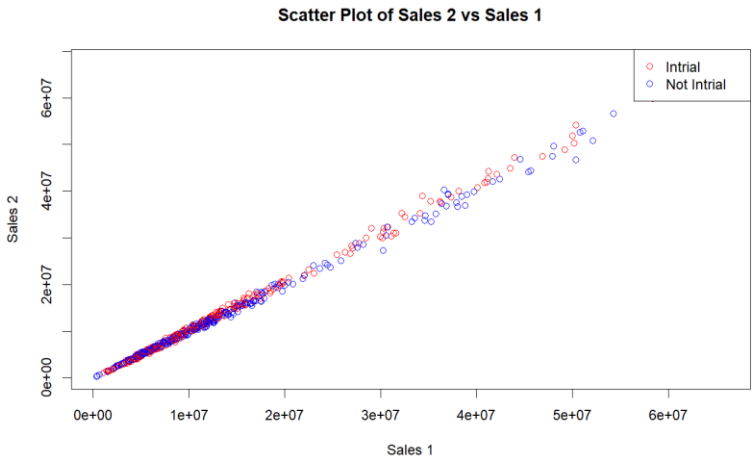


Figure 2: Scatterplot of Sales 2 against Sales 1, strong positive linear relationship observed

From the scatterplot in Figure 2 we can see a clear positive linear relationship between sales 1 and sales 2, so stores with higher sales before the change generally tend to have higher sales after the change, regardless of whether they were in the trial. At first glance, it might seem like the red points (intrial stores) are slightly higher than the blue points (non-intrial stores) for the same sales 1 value. This *could* suggest a potential positive effect of the new layout and signage. However, it's not a very strong visual difference, and we need the regression analysis to confirm this. The intrial variable must be in the regression model.

Stores vary in their baseline sales due to factors like location, size, customer demographics, etc. sales 1 captures these inherent differences. By including sales 1 in the regression model, we can account for these baseline differences, to allow more accurate isolation of the effect of the intrial variable (the new layout and signage) on sales_2.

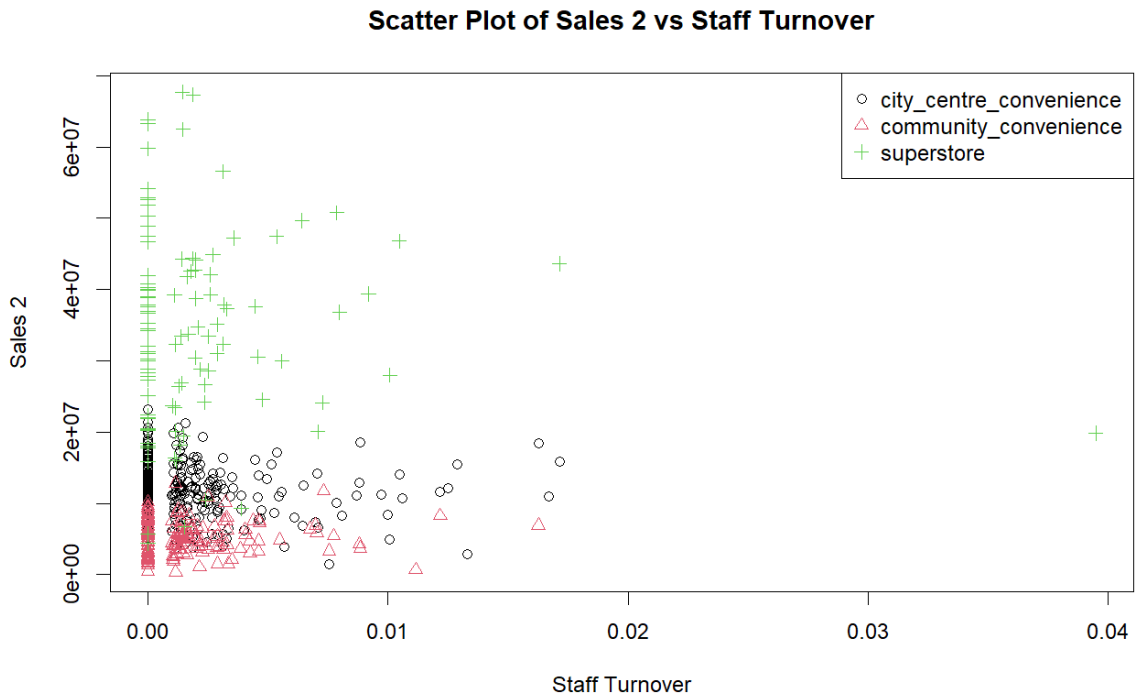


Figure 3: Scatterplot of Sales 2 against staff turnover, no apparent relationship

This scatter plot in Figure 3 shows the relationship between sales 2 (sales after the store layout and signage change) and staff turnover (proportion of staff who left during the data period). Different store types (outlettype) are represented by different shapes and colours. There isn't a strong, visually apparent relationship between sales 2 and staff turnover. The points seem somewhat randomly scattered across the plot, suggesting that staff turnover might not be a major driver of sales in a simple linear way, although it could have interaction with other variables like outlettype.

Another takeaway from Figure 3 is that there may be subtle differences in how staff turnover relates to sales for different store types. For example, city-centre-convenience stores (circles) with very low staff turnover seem to have the highest sales. However, this is just a visual observation and needs further investigation.

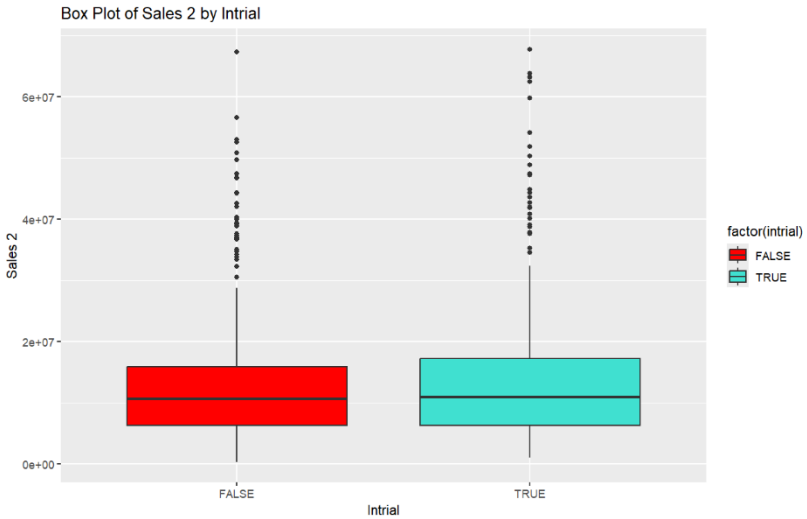


Figure 4: Boxplot of Sales 2 by Intrial, slightly higher median for Intrial = TRUE

From figure 4 we can see that the median sales 2 for stores in the trial (intrial = TRUE, turquoise box) appears to be slightly higher than the median for stores not in the trial (intrial = FALSE, red box). This suggests that the new store layout and signage *might* have had a positive effect on sales. The interquartile range, represented by the box height, is similar for both groups, indicating the variability in sales is comparable between trial and non-trial stores.

Both groups have a number of outliers, especially the trial group. These data points fall significantly outside the overall pattern of the data. The outliers could be due to various reasons (data errors, unusual store circumstances) and should be investigated further.

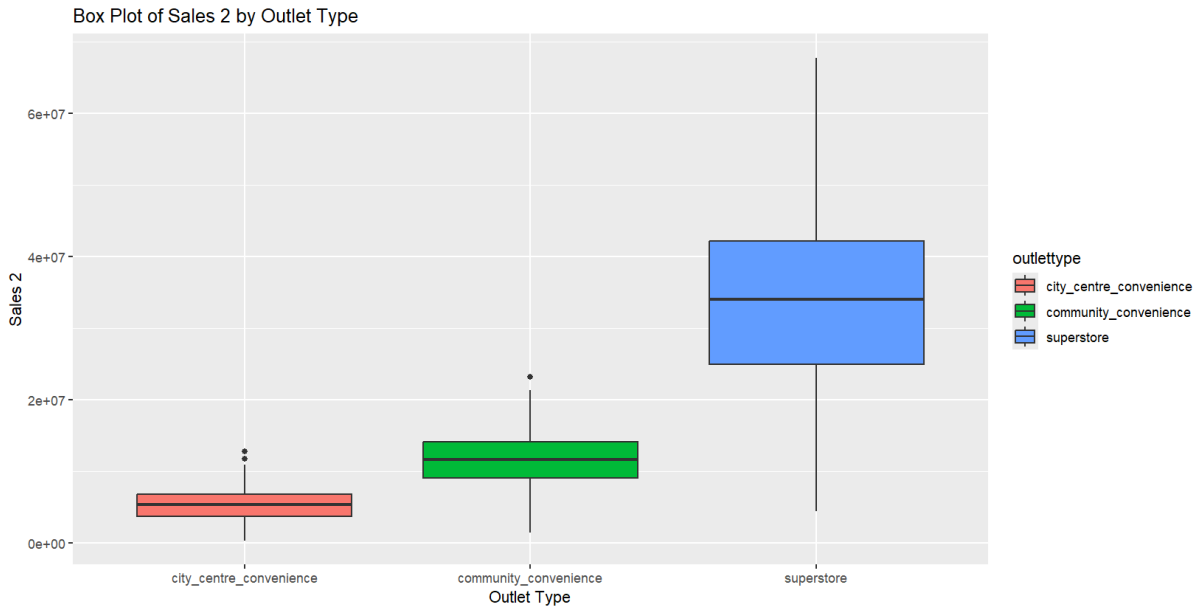


Figure 5: Boxplot of sales 2 by outlet type, apparent association

Figure 5's boxplot visually displays the distribution of sales 2 across different outlet type categories. Superstores have the highest median sales, indicating they generally achieve the highest sales figures after the change, followed by community stores, and lastly city-centre-convenience. This clearly shows that the outlet type is related to sales performance, even after the layout and signage change and thus should be included in the regression to account for inherent differences between store types such as size and product range.

Statistical Analysis:

I have chosen to utilise the multiple linear regression model to assess the impact of the new store layout and signage on sales. The model allows for the estimation of the effect of the layout and signage change while simultaneously controlling for other potential influences on sales, such as store type and staff turnover. It is important to do so because of the presence of potential confounding variables, such as store type and staff turnover, which could mask or distort the true effect of the layout and signage change.

Moreover, the linear regression model is a well-established and widely used statistical method that is relatively easy to interpret and understand. The results of the model, such as the coefficients and their associated p-values, can provide clear insights into the direction and magnitude of the relationships between the variables. The model takes log_sales_2 as the response variable and the following predictor variables:

- log_sales_1: Log of sales before the change.
- intrial: Whether the store was included in the trial (TRUE/FALSE).
- outlettype: The type of store (categorical variable).
- staff_turnover: The proportion of staff who left during the peri

```
Call:
lm(formula = log_sales_2 ~ log_sales_1 + intrial + outlettype +
    staff_turnover, data = sales_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.116915 -0.029729  0.002713  0.029308  0.118996

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -0.004669   0.061771  -0.076  0.93977
log_sales_1                     0.998639   0.003996 249.917  < 2e-16 ***
intrialTRUE                     0.009995   0.003801   2.630  0.00879 **
outlettypecommunity_convenience 0.040886   0.005339   7.658 8.95e-14 ***
outlettypesuperstore            0.043140   0.009116   4.732 2.84e-06 ***
staff_turnover                 -0.495051   0.585576  -0.845  0.39826
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04303 on 534 degrees of freedom
Multiple R-squared:  0.9972,    Adjusted R-squared:  0.9971
F-statistic: 3.744e+04 on 5 and 534 DF,  p-value: < 2.2e-16
```

Figure 6: Multiple linear regression output

In Figure 6 the five-number summary of residuals (min, 1st quartile, median, 3rd quartile, max) conveys the spread and central tendency of the errors. The median being close to zero is generally a good sign, suggesting that the model's predictions are, on average, unbiased.

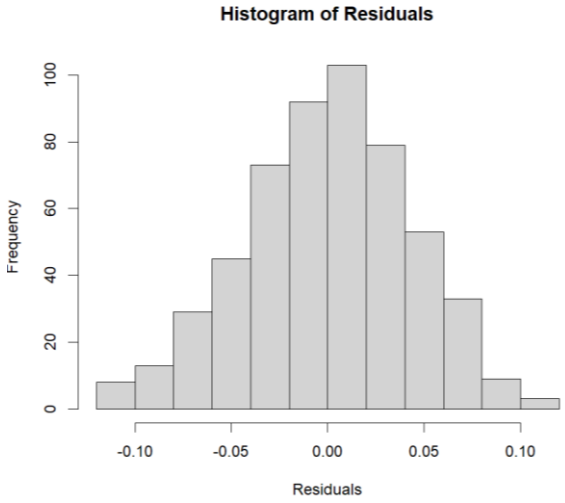


Figure 7: Histogram of residuals, approximately normal

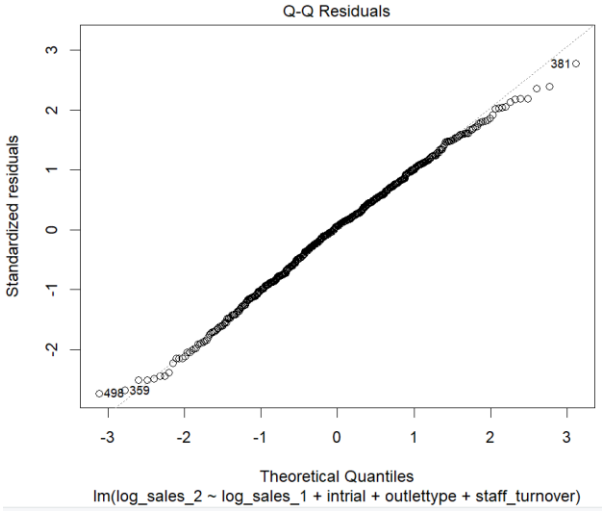


Figure 8: Normal Q-Q plot for residuals

Looking at the distribution of the residuals from the histogram in Figure 7 we can assume that the residuals approximate the normal. This is further reinforced by the points on the Q-Q plot in Figure 8 mostly following the straight diagonal line, so we can assert the normality of residuals.

The Residual standard error (0.04303) in Figure 6 is a measure of the typical size of the residuals, indicating how much the observed values deviate from the model's predictions on average. A lower residual standard error suggests a better fit, as the model's predictions are closer to the actual values.

The R-squared value (0.9972) indicates the model explains approximately 99.7% of the variability in the logarithm of sales_2. This is a very high value, suggesting a good overall fit, meaning the predictors are explaining the majority of variation in sales after the change. The adjusted R-squared (0.9971) is similar to the R-squared but takes into account the number of predictors in the model. It penalizes the inclusion of unnecessary predictors. In this case, the high adjusted R-squared further supports the strong fit of the model.

The F-statistic and its associated p-value test the overall significance of the model. The very low p-value (< 2.2e-16) indicates the model is statistically significant overall. This means that at least one of the predictors in the model is useful in explaining the variation in the logarithm of sales_2.

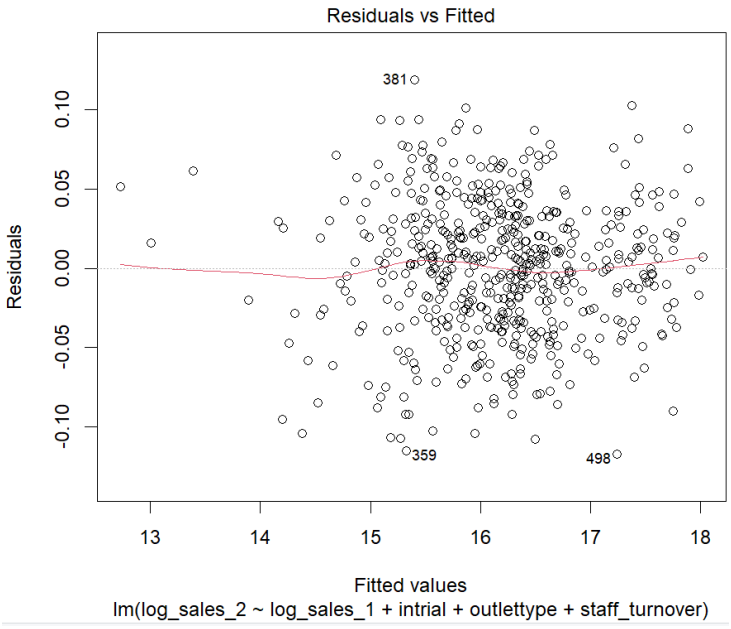
The Intercept -0.004669 is the estimated value of the logarithm of sales_2 when all other predictors are zero. However, with the log transformation and the presence of other predictors, the intercept might not have a meaningful practical interpretation in this context.

Again, in Figure 6 “intrialTRUE” (0.009995) is the coefficient (for the intrial variable when it's TRUE (i.e., the store was part of the trial). The positive coefficient suggests that being in the trial is associated with a slight increase in sales_2, even after accounting for other factors.

For outlettypes there is outlettypecommunity_convenience (0.040886) and outlettype-superstore (0.043140), these coefficients represent the differences in the logarithm of sales_2 between the respective store types (community_convenience and superstore) and the baseline store type (city_centre_convenience, which is omitted as the reference category). These coefficients indicate that different store types have different baseline sales levels, even after controlling for other factors.

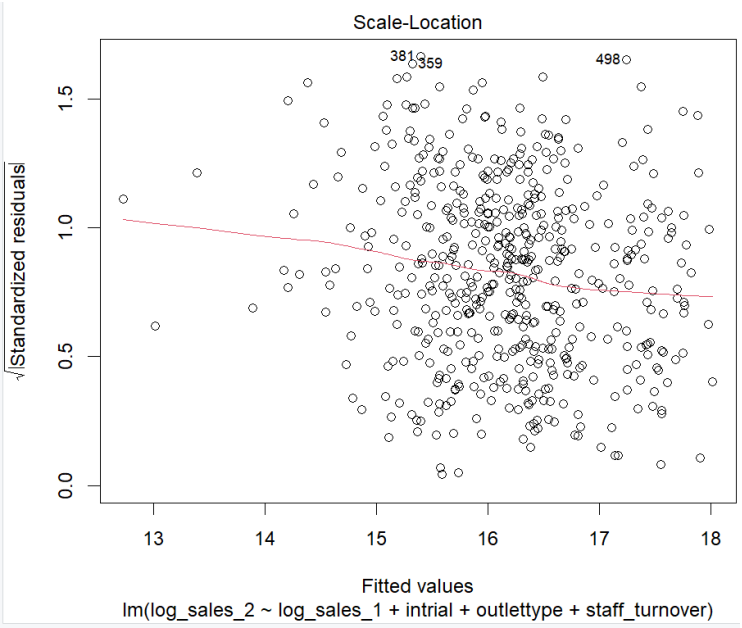
Lastly there is staff_turnover (-0.495051), the negative coefficient suggests that higher staff turnover might be associated with slightly lower sales.

The p-values in the Pr(>|t|) column indicates the statistical significance of each predictor. A low p-value (typically less than 0.05) suggests that the predictor is statistically significant, so it has a significant impact on the response variable. In this output, log_sales_1, intrialTRUE, outlettypecommunity_convenience, and outlettypesuperstore all have p-values less than 0.001, indicating they are highly significant predictors of sales_2. Only the staff_turnover variable has a quite a large p-value of 0.39826, therefore it is not a significant predictor.



<- Figure 9: Residuals vs Fitted values plot

In Figure 9 there appears to be a random scatter of points around the horizontal line at zero, and there is no funnelling of residuals, so we can assume there is a linear relationship between predictors and response, also the constant variance assumption is met.



From the scale-location plot in Figure 10, similarly to Figure 9 we see no clear pattern of points, and the red line point is relatively flat, indicating that the square root of the standardized residuals is fairly constant across the range of fitted values. This scale-location plot suggests that the homoscedasticity assumption is reasonably met. There's no strong evidence of a systematic pattern in the spread of the residuals, which would indicate heteroscedasticity (non-constant variance).

Figure 10: Scale-Location plot of Standardized residuals vs Fitted values

Since homoscedasticity is relatively present, we can assume the estimated standard errors of regression coefficients in Figure 6 are unbiased and consistent, which is necessary for making valid inferences about the significance of predictors using a t-test and significance of the overall model with the F-test. Homoscedasticity also means the ordinary least squares (OLS) estimator is the most efficient unbiased estimator.

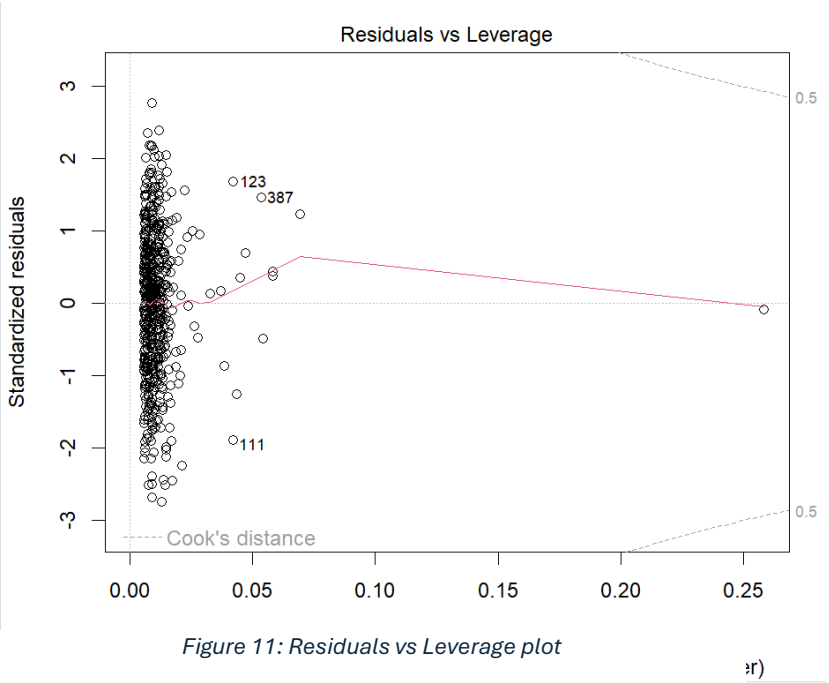


Figure 11: Residuals vs Leverage plot

In Figure 11 residuals are plotted against leverage, where leverage is a measure of how much influence a data point has on the estimated regression coefficients. Points with higher leverage have predictor values that are far from the mean of the predictor values, giving them more potential to pull the regression line towards them. Most points have low leverage, clustered on the left side of the graph showing that most of the data points have predictor values that are relatively close to the average. However, there are a few points towards the right with higher leverage, suggesting these data points have predictor values that are more unusual compared to the rest of the data. There are several points with large, standardized residuals making it clear outliers are present. No points have Cook's distance greater than 1 so there are no highly influential points in the data.

```
        Durbin-Watson test

data:  lm1
DW = 2.0351, p-value = 0.6264
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 12: Durbin-Watson test, checking autocorrelation

Finally, Figure 12 is the result of a Durbin-Watson test on the linear regression model I used. A DW statistic ~ 2 is ideal, 2.0351 is very close paired with a high p-value meaning the null hypothesis cannot be rejected that is, true autocorrelation is equal to 0. Then we can assume independence of residuals.

Conclusion:

The linear regression model indicates a potential positive effect of the new store layout and signage on sales. The coefficient for intrialTRUE was positive and statistically significant, suggesting that stores included in the trial experienced an increase in sales compared to those that were not, after controlling for other factors.

Regarding the fit of the model, it explained a very high proportion of the variability in sales. This suggests that the chosen predictors, including the log of sales before the change (log_sales_1), store type (outlettype), and the trial indicator (intrial), effectively capture the key factors influencing sales performance.

Despite the diagnostics showing relatively favourable results for the model there are still limitations of the analysis itself, selection bias may be present even though it has been mitigated by random sampling, as trial stores tended to have higher sales even before the change, mean of sales_1 for intrial = FALSE: 13874490.69 vs mean of sales_1 for intrial = TRUE: 14429702.81. Also, results of this analysis might not be generalizable to other retailers, store types, or time periods. The data was collected from a specific retailer during a particular period, and the results might vary under different circumstances.

The diagnostic plots indicated the presence of potential outliers in the data. While these outliers did not appear to be overly influential, it is important to acknowledge their potential impact on the results. Finally, there is an inherent limitation in the linear regression due to its assumption of linear relationships. While the diagnostic plots did not reveal strong evidence of non-linearity, it is possible that more complex relationships exist that the model does not capture.

In conclusion, the analysis provides evidence suggesting that the new store layout and signage had a positive effect on average sales. However, it is essential to consider the limitations of the analysis and acknowledge the potential influence of other factors, such as store type and potential selection bias.

Ayush Pradhan

(R code):

```
library(ggplot2)
library(fBasics)
library(car)
library(MASS)
library(lmtest)

# Read the data in
sales_data <- read.csv("sales_data.csv")
attach(sales_data)
head(sales_data)
tail(sales_data)
dim(sales_data)

# ---Univariate Analysis---
# Histograms of both sales_1 and sales_2
par(mfrow=c(1,2))
hist(sales_1, main = "Histogram of Sales 1", xlab = "Sales 1")
hist(sales_2, main = "Histogram of Sales 2", xlab = "Sales 2")

# Density plots of sales_1 and sales_2
par(mfrow=c(1,2))
plot(density(sales_1), main = "Density Plot of Sales 1", xlab = "Sales 1")
plot(density(sales_2), main = "Density Plot of Sales 2", xlab = "Sales 2")

# Summary statistics for sales_1 and sales_2
summary(sales_data[,c("sales_1","sales_2")])

# Standard deviation and IQR for sales_1 and sales_2
apply(sales_data[,3:4],2,sd)
apply(sales_data[,3:4],2,IQR)
```

Ayush Pradhan

```
# ---Bivariate Analysis---
# Scatter plot for sales_2 against sales_1
par(mfrow=c(1,1))
plot(sales_1, sales_2,
    main = "Scatter Plot of Sales 2 vs Sales 1",
    xlab = "Sales 1", ylab = "Sales 2",
    col = ifelse(sales_data$intrial == TRUE, "red", "blue"))  # Color points by intrial)
legend("topright", legend = c("Intrial", "Not Intrial"), col = c("red", "blue"), pch = 1)

# Scatter plots for sales_1 and sales_2 against staff_turnover
outlet_code <- rep(NA,length(outlettype))
outlet_code[which(outlettype=="community_convenience")] <- 1
outlet_code[which(outlettype=="city_centre_convenience")] <- 2
outlet_code[which(outlettype=="superstore")] <- 3

par(mfrow=c(1,1))
plot(staff_turnover, sales_1, col = outlet_code, pch = outlet_code,
    main = "Scatter Plot of Sales 1 vs Staff Turnover",
    xlab = "Staff Turnover", ylab = "Sales 1")
legend("topright", legend = levels(as.factor(outlettype)), col = 1:3, pch = 1:3)

plot(staff_turnover, sales_2, col = outlet_code, pch = outlet_code,
    main = "Scatter Plot of Sales 2 vs Staff Turnover",
    xlab = "Staff Turnover", ylab = "Sales 2")
legend("topright", legend = levels(as.factor(outlettype)), col = 1:3, pch = 1:3)

# Box plots for sales_1 and sales_2 by intrial
ggplot(sales_data, aes(x = factor(intrial), y = sales_1, fill = factor(intrial))) +
  geom_boxplot() +
  labs(title = "Box Plot of Sales 1 by Intrial", x = "Intrial", y = "Sales 1") +
  scale_fill_manual(values = c("FALSE" = "red", "TRUE" = "turquoise"))  # Assign colors
```

```
ggplot(sales_data, aes(x = factor(intrial), y = sales_2, fill = factor(intrial))) +
  geom_boxplot() +
  labs(title = "Box Plot of Sales 2 by Intrial", x = "Intrial", y = "Sales 2") +
  scale_fill_manual(values = c("FALSE" = "red", "TRUE" = "turquoise"))  # Assign colors

# Box plots for sales_1 and sales_2 by outlettype
ggplot(sales_data, aes(x = outlettype, y = sales_1, fill = outlettype)) +
  geom_boxplot() +
  labs(title = "Box Plot of Sales 1 by Outlet Type", x = "Outlet Type", y = "Sales 1")

ggplot(sales_data, aes(x = outlettype, y = sales_2, fill = outlettype)) +
  geom_boxplot() +
  labs(title = "Box Plot of Sales 2 by Outlet Type", x = "Outlet Type", y = "Sales 2")

# Paired t-test
t.test(sales_2, sales_1, paired = TRUE)

# ---Multiple linear regression model---
# Convert intrial and outlettype to factor variables (categorical)
intrial <- as.factor(intrial)
outlettype <- as.factor(outlettype)

# Log transformation of sales_1 and sales_2 due to right skewness
log_sales_1 <- log(sales_1)
log_sales_2 <- log(sales_2)

# Fit the linear regression model
lm1 <- lm(log_sales_2 ~ log_sales_1 + intrial + outlettype + staff_turnover, data = sales_data)
summary(lm1)

# --Diagnostic Checks—
```

Ayush Pradhan

```
vif(lm1) # Variance inflation factor
anova(lm1) # ANOVA test for the overall model (F-test)
dwtest(lm1) # Durbin-Watson test, check for autocorrelation (independence in residuals)

# Residuals vs Fitted Values Plot
plot(lm1, which = 1) # Check for linearity and constant variance

# Normal Q-Q Plot
plot(lm1, which = 2) # Checks for normality of residuals

# Scale-Location Plot
plot(lm1, which = 3) # Check for homoscedasticity

# Residuals vs Leverage Plot
plot(lm1, which = 5) # Identify influential points

# Histogram of Residuals
hist(resid(lm1), main = "Histogram of Residuals", xlab = "Residuals") # Distribution of residuals
```