

Freedom Mobile Online Training Analysis

Alexei Marcilio

May 1, 2017

Contents

Introduction	3
Business objectives.	3
Exploratory Data Analysis	3
Data Visualization	5
Hypotheses	9
Check of statistical conditions	9
Statistical Tests	11
Conclusions	15
Appendix	17

Introduction

Freedom Mobile is a Canadian wireless telecommunications provider owned by Shaw Communications. With 1,052,758 active subscribers (as of the end of December 2016) in urban areas of Ontario, British Columbia and Alberta, it is Canada's fourth-largest mobile operator.

Founded in 2008 as Wind Mobile by the communications company Globalive, Freedom was one of several new mobile carriers launched in Canada in 2008 after a government initiative to encourage competition in the wireless sector alongside Mobilicity (later acquired by Rogers Communications) and Public Mobile (later acquired by Telus). It initially launched mobile data and voice services in the Greater Toronto and Hamilton Area, Ontario on December 16, 2009 and two days later in Calgary, Alberta. Since then, Southern Ontario has been the main target of network expansion: first with Ottawa in Q1 2011, and then with about half a dozen additional regions, the most recent being Brantford on July 3, 2014.

Training is important to the company and considered integral to their success. Every salesperson, store manager, and customer care employee are asked to complete online training courses each month. Online quizzes also accompany these training modules. The company would like to measure the success of their online training program.

Business objectives.

Freedom Mobile manages three different types of operations, Corporate Retail, Dealer Retail and Customer Care. Corporate Retail includes outlets that are run by head office, and Dealer Retail covers franchise operations. Customer Care provide support to customers and include three call centers, one in Windsor Ontario, Telepresence a call center in Cairo Egypt, and Startek, a center in Manila Philippines.

In order to better understand the impact and success of their online training program Freedom would like to have the following research questions answered:

1. *Is there a significant difference in completion rate and quiz scores between different channels?*
2. *Is there a significant difference in these scores between Customer Care's offshore and onshore resources?*
3. *Is there a significant difference among each Corporate Retail group, and between each Dealer Retail group?*
4. *Have the completion rates changed over time?*
5. *Are quiz scores correlated with completion rates?*

Exploratory Data Analysis

Table 1 shows the method in which the csv file was loaded into R studio. The file is comma delimited. There are seven variables, five of which are considered factors and two which are categorical. These include the employee id (id_number), channel, group_name, province and city. The two numeric variables are comp_rate (completion rate) and quiz (average quiz score). The comp_rate and quiz variables are continuous variables.

Table 1: R code to load freedom csv data.

```
setwd("/Users/alexeimarcilio/Desktop/portfolio/telecom")
getwd()
```

```
## [1] "/Users/alexeimarcilio/Desktop/portfolio/telecom"

freedom <- read.csv("freedom.csv")
```

We can examine the structure of the data. The id_numbers, channels, group_names, prov, and cities are considered factors, and comp_rate and quiz variables are considered numeric (Table 2). This shows the data is in the correct format for analysis.

Table 2: Structure of the freedom R data.

```
## 'data.frame': 2169 obs. of 7 variables:
## $ id_number : Factor w/ 2169 levels "aaamir","aadina",...: 1 2 3 4 5 6 7 8 9 10 ...
## $ channel : Factor w/ 3 levels "Care","Corp-retail",...: 2 2 2 2 2 1 2 2 1 2 ...
## $ group_name: Factor w/ 21 levels "CORPCA01-Calgary",...: 6 1 4 9 2 21 6 3 21 6 ...
## $ prov : Factor w/ 4 levels "", "Alberta", "BC",...: 4 2 4 3 2 1 4 4 1 4 ...
## $ city : Factor w/ 28 levels "", "Barrie", "Brampton",...: 13 6 23 24 7 1 13 18 1 13 ...
## $ comp_rate : num 0.981 0.968 1 0.957 0.857 1 0.935 0.917 0.95 0.982 ...
## $ quiz : num 92.2 98.8 92.5 90 92.1 ...
```

Table 3 shows a numerical summary for completion rate and quiz scores. We can see that the mean of the completion rate is 0.8683 and the median is .9090 which shows that the data is left skewed. The interquartile range is 0.131 and the values range from 0 to 1 (no activities completed - all activities completed)

For the quiz scores we can see that the mean is 90.63 and the median is 92.25 which also shows that the data is left skewed. The interquartile range is 8.57 and the values range from 10 to 100.

Table 3: Summary of the completion rate and quiz scores.

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000  0.8330  0.9090  0.8683  0.9640  1.0000

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 10.00   87.33   92.25   90.63   95.90  100.00     40
```

We can see that there are fairly similar numbers of records of the three channels (Table 4), with 768 observations for Care and 721 and 680 for Corporate-retail and Dealer-retail respectively.

Table 4: Frequencies of Channels.

```
##
##      Care  Corp-retail Dealer-retail
##      768      721      680
```

We plot the completion rates vs quiz scores in order to detect any potential outliers (Figure 1). We can see that all the completion rates and quiz scores seem reasonable and there are no visible outliers. We used color to differentiate each channel and it's interesting to note that there seems to be a pattern showing the Care channel to have lower completion rates and quiz scores. This is something we will explore later. The data seem concentrated above a .75 completion rate and quiz scores of over 75%.

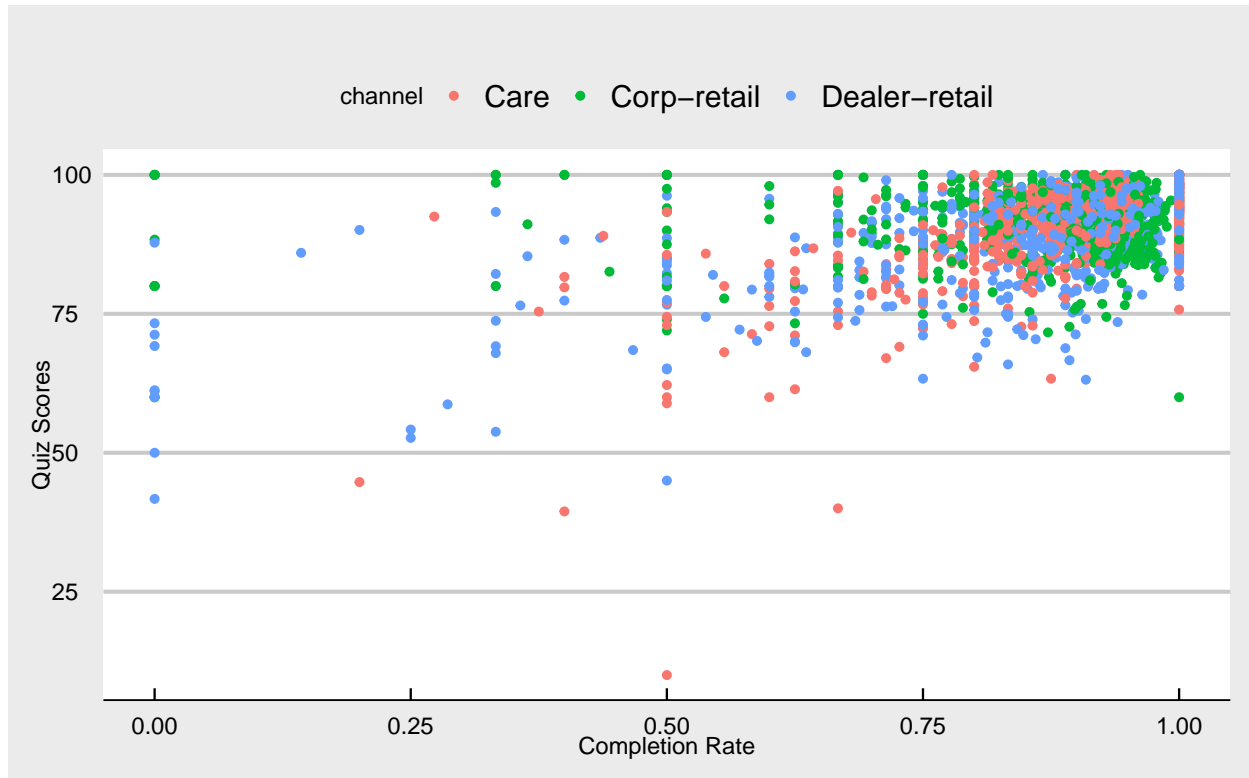


Figure 1: Plot of completion rates vs quiz scores.

Data Visualization

Let's example the distribution of completion rates and quiz scores.

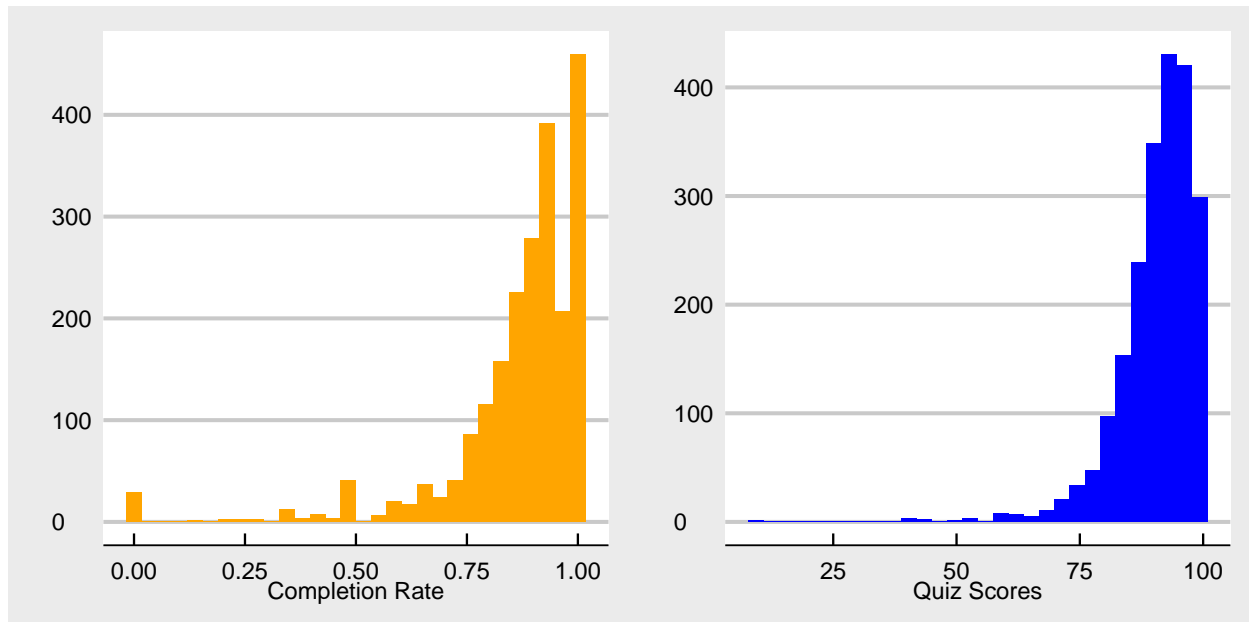


Figure 2: Distribution of completion rates and quiz scores.

Figure 2 shows the distribution of completion rates and quiz scores. We can see that both are unimodal and skewed to the left. This is understandable as there are many scores that are high. The mode for completion

rate is 1 and the mode for the quiz scores is 100. The distribution is fairly normal. There are quite a few 0 values for completion rates, in fact there are at 29 employees who have not completed any activities.

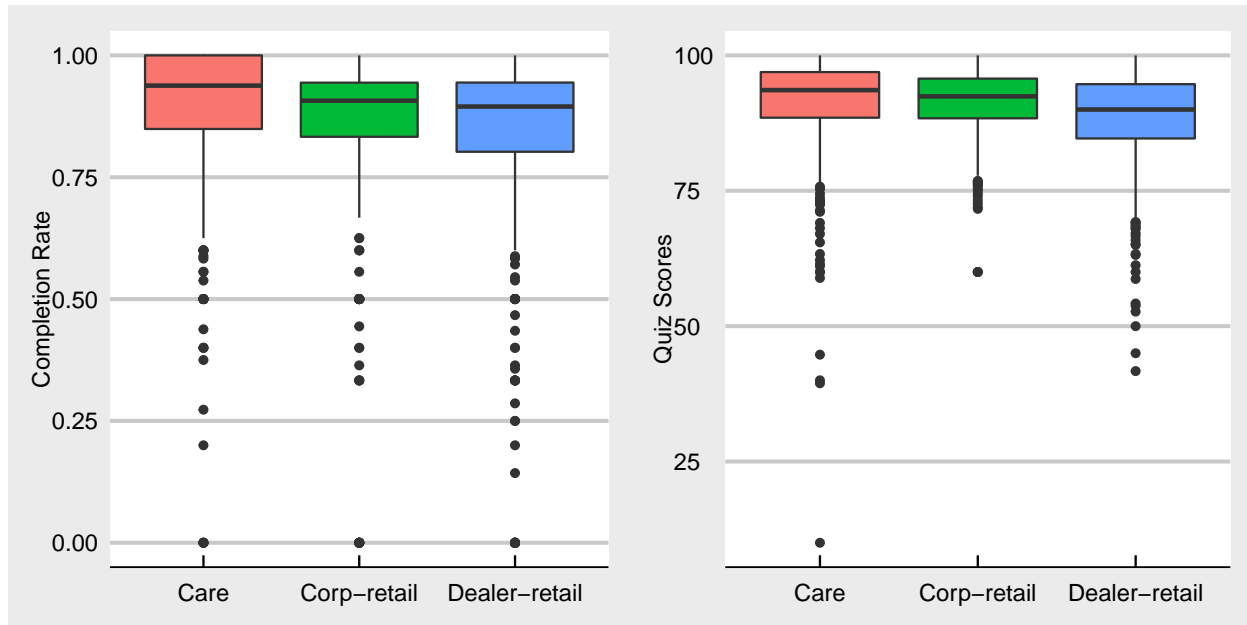


Figure 3: Box plots of completion rates & quiz scores by channel.

Box plots of completion rates and quiz scores show that the median values are fairly similar among the channels (Figure 3). In addition we can see that the data is left skewed for all channels and that the Care channel has a slightly higher median value than the Corporate Retail channel, which is higher than the Dealer Retail channel for both completion rates and quiz scores.

As the data is left skewed the median is a good measure of the distribution. The median value of completion rates for the Care channel is 0.938, which is higher than that of Corporate Retail (0.907) and Dealer Retail (0.895).

The median values for quiz scores are 93.59 for Care, 92.43 for Corporate Retail and 90 for Dealer Retail.

Let's look at the distribution of the offshore and onshore resources for the Care channel. Freedom operates call centers in Windsor Ontario, Cairo Egypt (Telepresence) and Manila Philippines (Startek). We can see that the median values are fairly close to each other however the Teleperformance group has the lowest median values for both completion rates and quiz scores (Figure 4). All the scores are left skewed and Teleperformance has the lowest values for completion rates and quiz scores as well as the largest interquartile range.

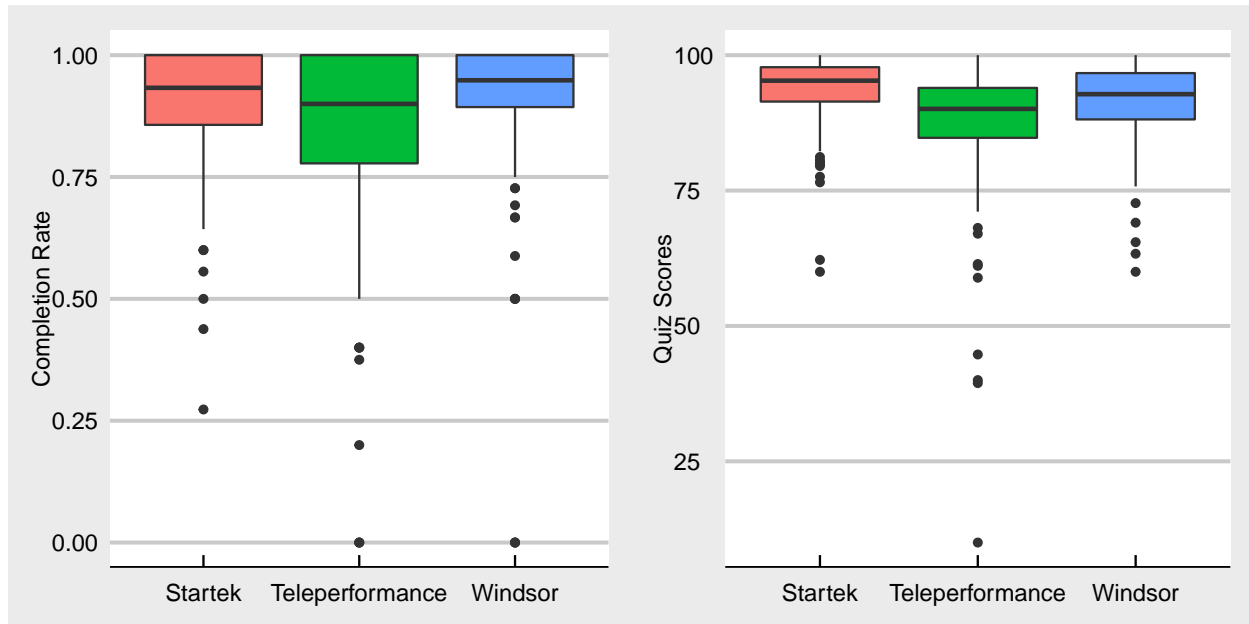


Figure 4: Box plots of completion rates & quiz scores by Care group.

Let's compare the completion rates for the Corporate Care groups. We can see in Figure 5 that the median values are somewhat similar, however the range of data differs between the groups. The completion rates show more outliers, all left skewed, for the Ontario groups than for the western groups.

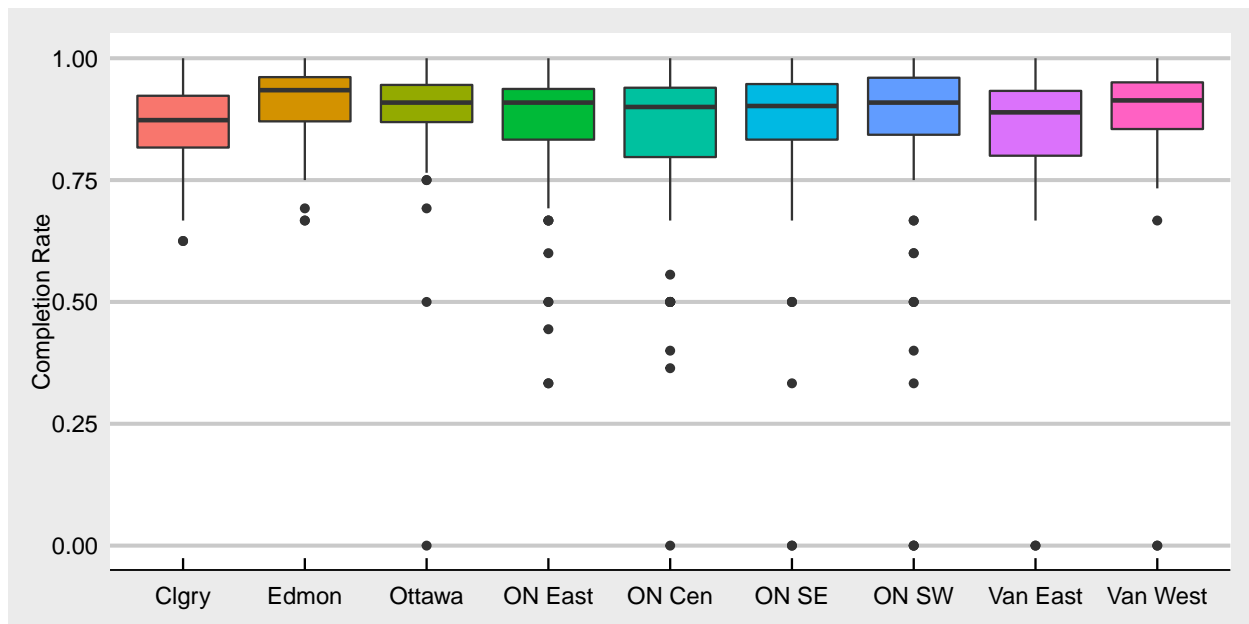


Figure 5: Box plots of completion rates & quiz scores by Corporate Retail group.

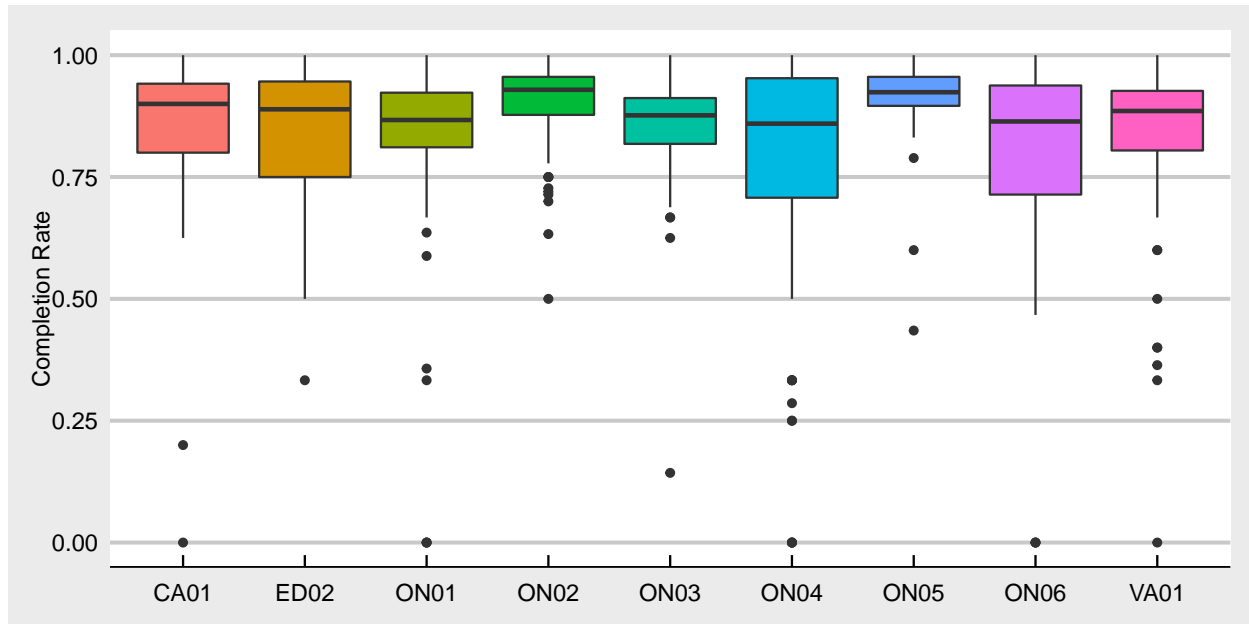


Figure 6: Box plots of completion rates & quiz scores by Dealer Retail group.

If we do the same for the Dealer Retail groups the box-plot does not reveal great differences between the medians, although the variance among the groups is large (Figure 6). Ontario-04 has a very narrow IQR compared to Ontario-04 for example.

If we now examine the completion rates over time there seems to be a general downward trend (Figure 7). Completion rates for the entire company were 1 until October of 2016 when some employees did not finish the learning activities. This trend seemed to accelerate as more and more employees did not complete activities in the ensuing months.

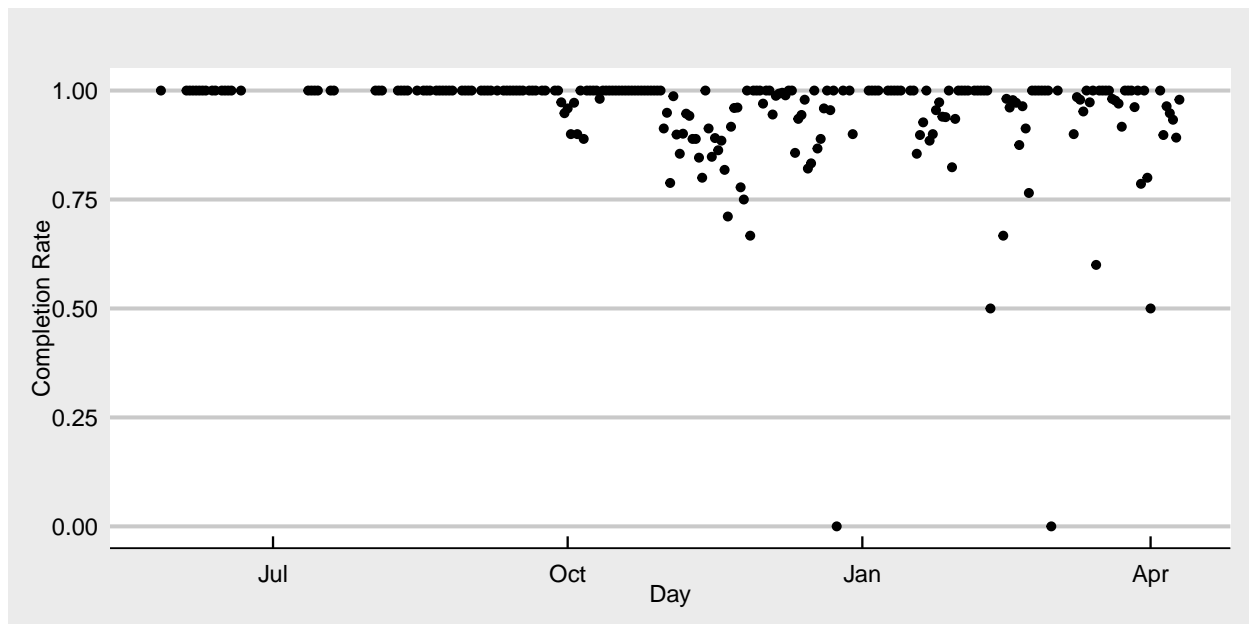


Figure 7: Change in the completion rates over time.

The other correlation we wish to examine is completion rates vs. quiz scores. Is there a correlation between these two variables? We plotted this in Figure 1. A correlation is not evident in this plot as the scores are concentrated among the higher values.

Hypotheses

Here are the hypotheses for each of our business questions:

1. Is there a significant difference in completion rate and quiz scores between different channels?__
 - i) $H_0 : \mu_{care} = \mu_{Corp} = \mu_{Retail}$
 - ii) $H_a : \mu_{care} \neq \mu_{Corp} \neq \mu_{Retail}$
2. Is there a significant difference in these scores between Customer Care's offshore and onshore resources?__
 - i) $H_0 : \mu_{Windsor} = \mu_{Startek} = \mu_{Teleperformance}$
 - ii) $H_a : \mu_{Windsor} \neq \mu_{Startek} \neq \mu_{Teleperformance}$
3. Is there a significant difference among each Corporate Retail group, and between each Dealer Retail group?__
 - *Corporate Retail*
 - $H_0 : \mu_{Corp-Calgary} = \mu_{Corp-Edmonton} = \dots = \mu_{Corp-VanWest}$
 - $H_a : \mu_{Corp-Calgary} \neq \mu_{Corp-Edmonton} \neq \dots \neq \mu_{Corp-VanWest}$
 - *Dealer Retail*__
 - $H_0 : \mu_{DEALON01} = \mu_{DEALON02} = \dots = \mu_{DEALVA01}$
 - $H_a : \mu_{DEALON01} \neq \mu_{DEALON02} \neq \dots \neq \mu_{DEALVA01}$
4. Have the completion rates changed over time?__
 - i) $H_0 : \beta_1 = 0$. *The true linear model has a slope of 0.*
 - ii) $H_0 : \beta_1 \neq 0$. *The true linear model does not has a slope of 0.*
5. Are quiz scores correlated with completion rates?__
 - i) $H_0 : \beta_1 = 0$. *The true linear model has a slope of 0.*
 - ii) $H_0 : \beta_1 \neq 0$. *The true linear model does not has a slope of 0.*

Check of statistical conditions

##ANOVAs

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

These conditions are satisfied. The data are independent as there is no collaboration or communication among the three channels of business, nor among each Corporate Retail group or Dealer Retail group. Also the side by side box plots shown in previous figures show somewhat similar distributions for these groups.

Table 5: Mean and STD values for channels, Care groups and Dealer Retail groups.

##	channel	mean	std
## 1	Care	0.8985352	0.1449097
## 2	Corp-retail	0.8601748	0.1552549
## 3	Dealer-retail	0.8427941	0.1793946

```
##           group      mean      std
## 1      Startek 0.9159107 0.1004143
## 2 Teleperformance 0.8501294 0.1934527
## 3      Windsor 0.9153545 0.1432597

##           group      mean      std
## 1 DEALCA01 0.8408205 0.19853501
## 2 DEALE02 0.8381622 0.15663750
## 3 DEALON01 0.8295773 0.18764543
## 4 DEALON02 0.9018929 0.09263271
## 5 DEALON03 0.8502143 0.12112981
## 6 DEALON04 0.7850323 0.23679345
## 7 DEALON05 0.9150879 0.07536823
## 8 DEALON06 0.7905833 0.23116611
## 9 DEALVA01 0.8397564 0.17788046
```

In addition if we example the mean and standard deviations for each channel and each Care group we can see that they are similar, which satisfies the condition that the variability across the groups is about equal (Table 5). However the standard deviations of the Dealer Retail groups are not that similar. They range from 0.07536823 to 0.23679345.

##Linear Models

To assess whether the linear model is reliable, we need to check for (1) linearity, (2) nearly normal residuals, and (3) constant variability.

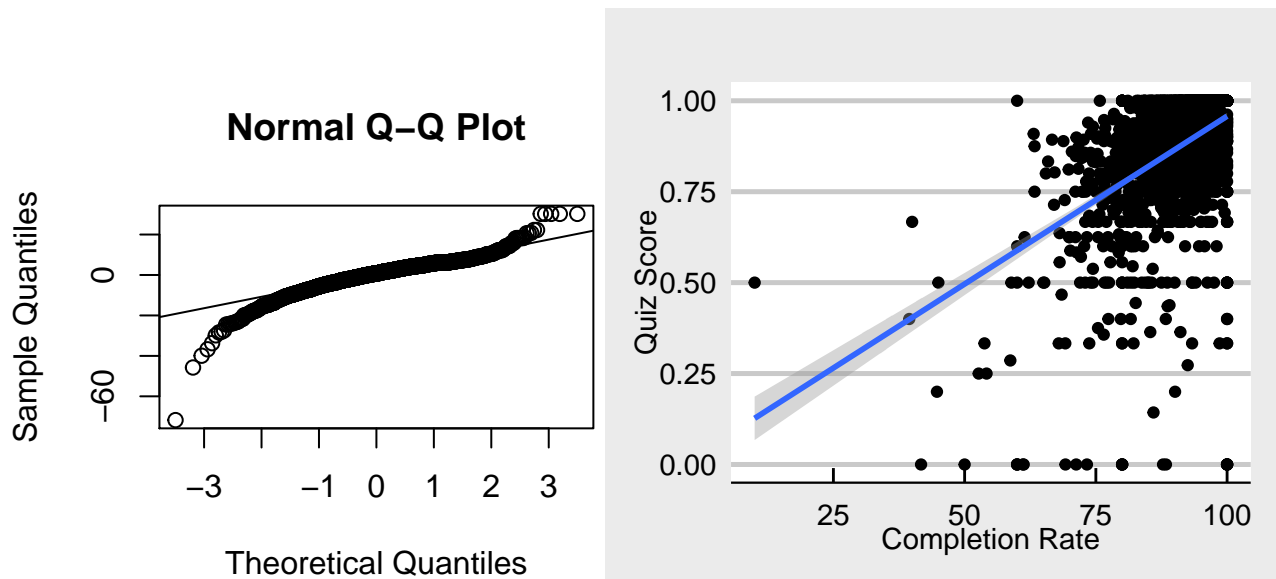


Figure 8: Normal probability plot of the residuals, and scatterplot with a fitted least squares line.

We can see from Figure 8 that the data satisfies the nearly normal condition. The residuals indicate skewing at each end of the plot, but these balance each other and we can state the data is nearly normal. Also the plot of the data show no discernible pattern so a linear model is a reasonable assumption. There is some indication that the data do not show constant variability at lower completion rates as we can see from the increasing width of the confidence interval. This is not such a large spread so we will proceed.

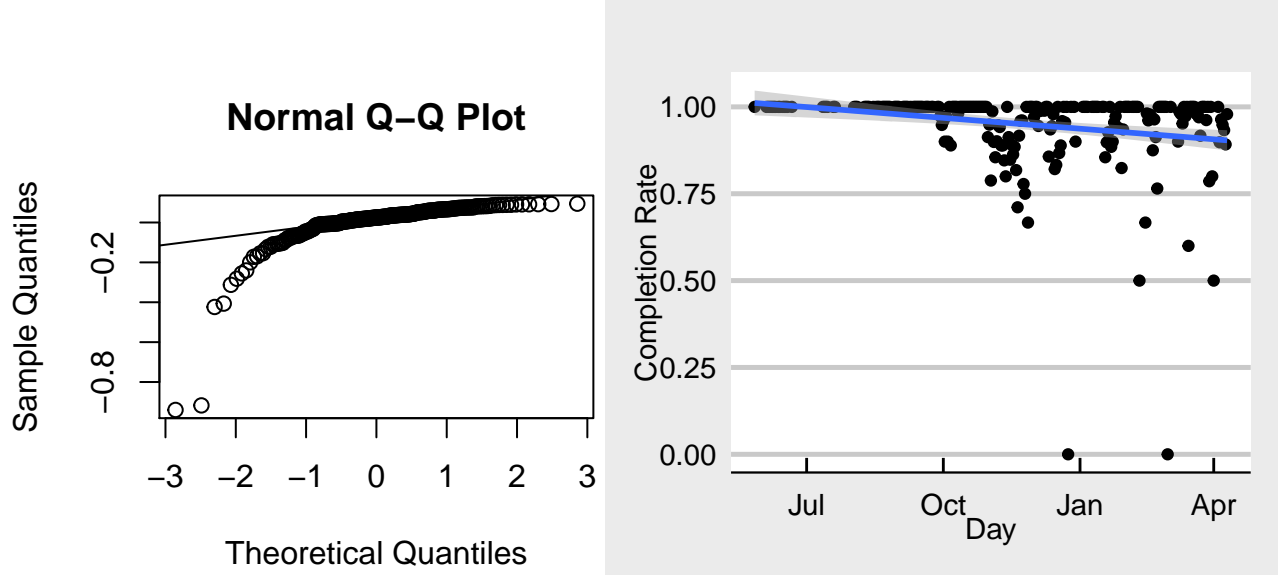


Figure 9: Normal probability plot of the residuals, and scatterplot with a fitted least squares line.

In contrast the plot of the residuals of the correlation between date and completion rates show a skewing at one end (Figure 9). Although the plot shows some linearity and almost constant variance, the fact that the data is so skewed precludes the normality assumption.

Statistical Tests

1._Is there a significant difference in completion rate and quiz scores between different channels?_

i) $H_0 : \mu_{care} = \mu_{Corp} = \mu_{Retail}$

ii) $H_a : \mu_{care} \neq \mu_{Corp} \neq \mu_{Retail}$

Table 6: ANOVA for completion rate each channel

```
##               Df Sum Sq Mean Sq F value    Pr(>F)
## channel         2   1.19   0.5960    23.34 9.36e-11 ***
## Residuals      2166   55.31   0.0255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the results in Table 6 we can reject the null hypothesis and conclude that at least one of the channels has a different mean completion rate than the others.

2._Is there a significant difference in these scores between Customer Care's offshore and onshore resources?_

i) $H_0 : \mu_{Windsor} = \mu_{Startek} = \mu_{Teleperformance}$

ii) $H_a : \mu_{Windsor} \neq \mu_{Startek} \neq \mu_{Teleperformance}$

Table 7: ANOVA for completion rate each care organization.

```
##               Df Sum Sq Mean Sq F value    Pr(>F)
## group_name     2   0.638   0.3190    15.78 1.93e-07 ***
## Residuals      765  15.468   0.0202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the results in Table 7 we can reject the null hypothesis and conclude that at least one of the care groups has a different mean completion rate than the others.

3. Is there a significant difference among each Corporate Retail group, and between each Dealer Retail group?_

- *Corporate Retail*

- $H_0 : \mu_{Corp-Calgary} = \mu_{Corp-Edmonton} = \dots = \mu_{Corp-VanWest}$

- $H_a : \mu_{Corp-Calgary} \neq \mu_{Corp-Edmonton} \neq \dots \neq \mu_{Corp-VanWest}$

Table 8: ANOVA for completion rate each Corporate Retail Group.

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## group_name  8  0.202  0.02530    1.05  0.397
## Residuals 712 17.153  0.02409
```

Based on the results in Table 8 we fail to reject the null hypothesis and conclude that there's no significant difference in the mean completion rates between Corporate Retail Groups.

- *Dealer Retail*_

- $H_0 : \mu_{DEALON01} = \mu_{DEALON02} = \dots = \mu_{DEALVA01}$

- $H_a : \mu_{DEALON01} \neq \mu_{DEALON02} \neq \dots \neq \mu_{DEALVA01}$

Table 9: ANOVA for completion rate each Dealer Retail Group.

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## group_name  8  1.369  0.17109    5.605 6.78e-07 ***
## Residuals 671 20.483  0.03053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the results in Table 9 we can reject the null hypothesis and conclude that at least one of the Dealer Retail Groups has a different mean completion rate than the others.

4. Have the completion rates changed over time?_

i) $H_0 : \beta_1 = 0$. The true linear model has a slope of 0.

ii) $H_0 : \beta_1 \neq 0$. The true linear model does not has a slope of 0.

Not performed as data did not conform to normality assumptions.

5. Are quiz scores correlated with completion rates?_

i) $H_0 : \beta_1 = 0$. The true linear model has a slope of 0.

ii) $H_0 : \beta_1 \neq 0$. The true linear model does not has a slope of 0.

Table 10: Regression analysis for completion rates vs. quiz scores.

```
##
## Call:
## lm(formula = quiz ~ comp_rate, data = freedom)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -71.750  -3.280   0.802   4.347  30.212
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.7876    0.8618   80.98  <2e-16 ***
## comp_rate    23.9250    0.9744   24.55  <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.859 on 2127 degrees of freedom
## (40 observations deleted due to missingness)
## Multiple R-squared:  0.2209, Adjusted R-squared:  0.2205
## F-statistic: 602.9 on 1 and 2127 DF,  p-value: < 2.2e-16
```

Table 10 shows there is a significant linear relationship between completion rates and quiz scores as the p-value ~ 0 .

The least squares regression line for the linear model is:

$\hat{y} = 69.79 + 23.93 * comp_rate$. For every 10% increase in completion rate we can expect a 72.183 increase in quiz scores.

$R^2 = 0.2209$ which indicates there's a weak correlation between the variables. Only 22.09 % of the variation in quiz scores can be explained by completion rate.

##Additional Tests

As some of the ANOVA tests were significant we will run pairwise t-tests using the Bonferroni correction to determine which groups are actually different from one another.

Table 11: pairwise t-tests with a Bonferroni correction for completion rates vs. channel.

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  comp_rate and channel
##
##           Care      Corp-retail
## Corp-retail 1.2e-05 -
## Dealer-retail 1.3e-10 0.13
##
## P value adjustment method: bonferroni
```

Table 11 shows that Care has a significantly higher completion rate from Corporate Retail and Dealer Retail which are not significantly different. The mean Care completion rate is about 90% compared to Corporate Retail at 86% and Dealer Retail at 84%.

Table 12 shows that among the Care groups, Teleperformance has a significantly lower mean completion rate at 85%. The other two groups show no significant difference.

Table 12: pairwise t-tests with a Bonferroni correction for completion rates among Care groups.

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  comp_rate and group_name
##
##           Startek Teleperformance
## Teleperformance 7.0e-07 -
## Windsor          1          9.2e-06
##
## P value adjustment method: bonferroni
```

Table 13: pairwise t-tests with a Bonferroni correction for completion among Dealer Retail groups.

```
##
## Pairwise comparisons using t tests with pooled SD
##
```

```
## data: comp_rate and group_name
##
##      DEALCA01 DEALEDO2 DEALON01 DEALON02 DEALON03 DEALON04 DEALON05
## DEALEDO2 1.00000 - - - - -
## DEALON01 1.00000 1.00000 - - - -
## DEALON02 1.00000 1.00000 0.20300 - - -
## DEALON03 1.00000 1.00000 1.00000 1.00000 - -
## DEALON04 1.00000 1.00000 1.00000 9.7e-05 0.46147 - -
## DEALON05 0.96070 0.87326 0.03033 1.00000 0.71312 3.5e-06 -
## DEALON06 1.00000 1.00000 1.00000 0.00642 1.00000 1.00000 0.00075
## DEALVA01 1.00000 1.00000 1.00000 0.86522 1.00000 1.00000 0.19262
##      DEALON06
## DEALEDO2 -
## DEALON01 -
## DEALON02 -
## DEALON03 -
## DEALON04 -
## DEALON05 -
## DEALON06 -
## DEALVA01 1.00000
##
## P value adjustment method: bonferroni
```

The pairwise tests of Dealers (Table 13) are inconclusive. Group 4 is significantly less than 5 or group 2 but it's difficult to draw conclusions here.

##Confidence Intervals

We can also show that there's a significant different between the completion rate for each channel by considering the confidence intervals.

Table 14: Confidence intervals for the mean completion rates of each channel

```
## [1] "Care channel"
## Single mean
## Summary statistics:
## mean = 0.8985 ; sd = 0.1449 ; n = 768
## Standard error = 0.0052
## 95 % Confidence interval = ( 0.8883 , 0.9088 )
## [1] "Corporate Retail channel"
## Single mean
## Summary statistics:
## mean = 0.8602 ; sd = 0.1553 ; n = 721
## Standard error = 0.0058
## 95 % Confidence interval = ( 0.8488 , 0.8715 )
## [1] "Dealer Retail channel"
## Single mean
## Summary statistics:
## mean = 0.8428 ; sd = 0.1794 ; n = 680
## Standard error = 0.0069
## 95 % Confidence interval = ( 0.8293 , 0.8563 )
```

We can see from Table 14 that the Care channel has a significantly higher mean completion rate than the other channels by examining the confidence intervals. We are 95% confident that if $\mu_{CareChannel}$ (the average completion rate) were known, the interval between 0.8883 , 0.9088 would contain it. This interval does not overlap with the intervals of the other channels so we know it's statistically significantly higher than the rest. The confidence intervals for the Corporate Retail and Dealer Retail channels do overlap so we know that they are not different from each other.

Figure 10 show a comparison of the means of all these groups.

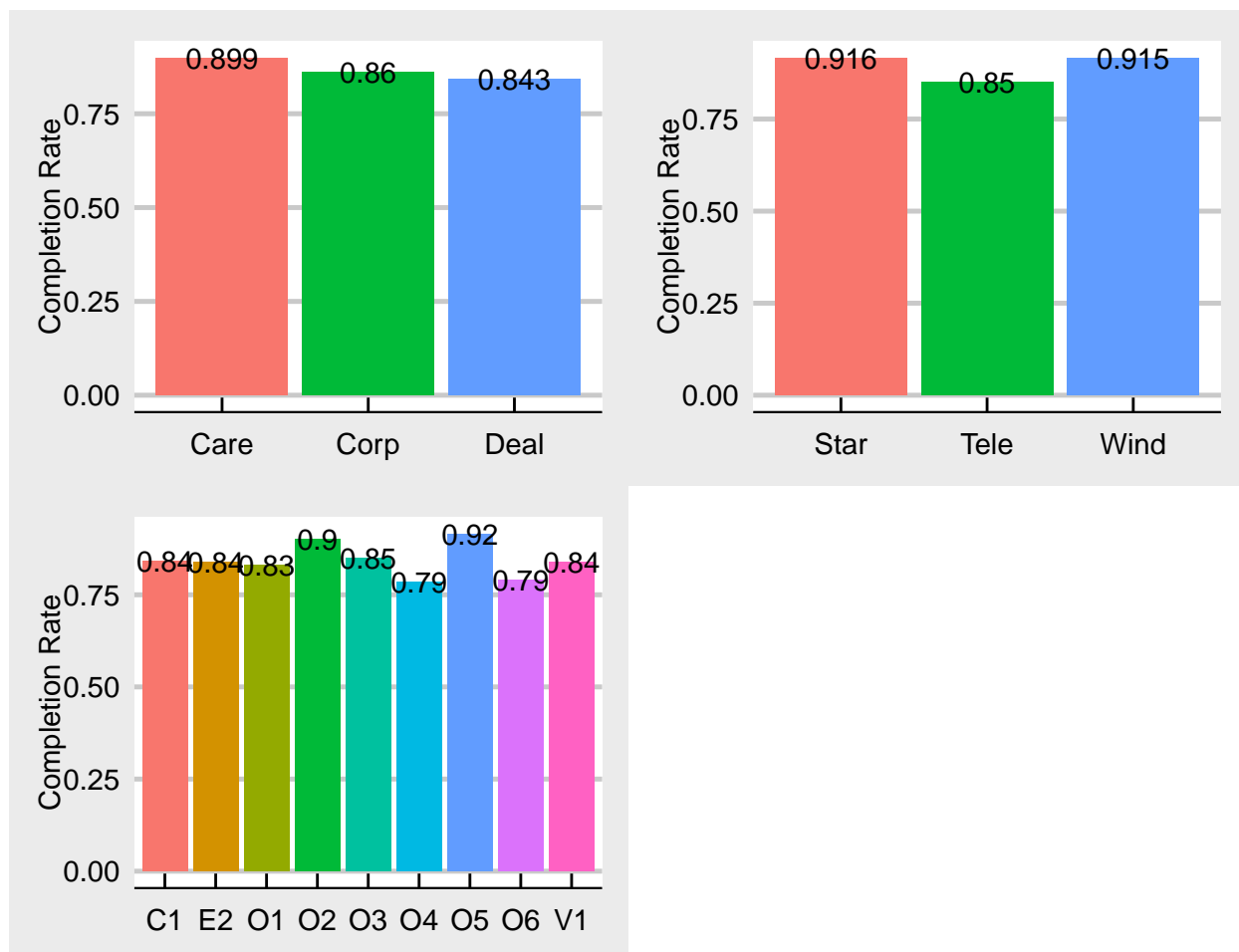


Figure 10: Comparison of mean completion rates between channels, and among Care and Dealer groups.

Conclusions

Freedom Mobile, the Canadian wireless telecommunications provider, wished to assess the success of their online training modules. Specifically they wanted to know if the completion rates differed among their three major channels (Corporate Retail, Dealer Retail and Customer Care) and if within these groups there were also differences. In addition they wanted to know if completion rates had changed over time, and if quiz scores were correlated to completion rates.

A statistical analysis was undertaken to find the answers to these questions. The results are as follows:

- There are indeed differences in the completion rates among the three channels.
 - Care has a significantly higher completion rate than that of the other groups.
- There are also significant differences within the three Care groups and within Dealer Retailers, but not among Corporate Retailers.

- The Teleperformance group has a significantly lower completion rate than the other Care groups.
- There is only a very weak correlation between completion rates and quiz scores.
- It could not be determined if completion rates changed over time as the data violated normality assumptions.

Further research would benefit Freedom Mobile. It would be useful to determine if completion rates relate to the financial performance of different groups. Also, the factors that contribute to the differences between groups warrants further study.

Appendix

R studio was used exclusively for the analyses in this report. The following is an extract of the markdown file used for this report.

```
require(ggplot2)
require(ggthemes)
require(captioner)
require(gridExtra)
require(ggmosaic)
require(lsr)

download.file("http://www.openintro.org/stat/data/nc.RData", destfile = "nc.RData")
load("nc.RData")

figs <- captioner(prefix="Figure")
tbls <- captioner(prefix="Table")

setwd("/Users/alexeimarcilio/Desktop/portfolio/telecom")
getwd()
freedom <- read.csv("freedom.csv")

str(freedom)

summary(freedom$comp_rate)
summary(freedom$quiz)

table(freedom$channel)

p <- qplot(freedom$comp_rate, freedom$quiz, data=freedom, xlab="Completion Rate", ylab="Quiz Scores", main="Freedom of the Press")
#plot(freedom$comp_rate, freedom$quiz, main = "", ylab = "Quiz Scores", xlab = "Completion Rate", cex = 1.5)

plot1 <- qplot(freedom$comp_rate,
  geom="histogram", fill=I("orange"),
  xlab = "Completion Rate") + theme_economist_white()

plot2 <- qplot(freedom$quiz,
  geom="histogram", fill=I("blue"),
  xlab = "Quiz Scores") + theme_economist_white()

grid.arrange(plot1, plot2, ncol=2)

Mode <- function(x) {
  if (is.numeric(x)) {
    x_table <- table(x)
    return(as.numeric(names(x_table)[which.max(x_table)]))
  }
}

box.c <- ggplot(freedom, aes(x=factor(channel), y=comp_rate))
```

```

box.c <- box.c + geom_boxplot(aes(fill = channel)) + theme_economist_white() + theme(legend.position="n
  theme(axis.title.x=element_blank()) + labs(y = "Completion Rate")

box.q <- ggplot(freedom, aes(x=factor(channel), y=quiz))

box.q <- box.q + geom_boxplot(aes(fill = channel)) + theme_economist_white() + theme(legend.position="n
  theme(axis.title.x=element_blank()) + labs(y = "Quiz Scores")

grid.arrange(box.c, box.q, ncol=2)
#pl3 <- pl2 + ggtitle("Weight by Sex") + xlab("Sex 1=M, 0=F")

freedom.care <- subset(freedom, freedom$channel=="Care")

care.c <- ggplot(freedom.care, aes(x=factor(group_name), y=comp_rate))

care.c <- care.c + geom_boxplot(aes(fill = group_name)) + theme_economist_white() + theme(legend.positi
  theme(axis.title.x=element_blank()) + labs(y = "Completion Rate")

care.q <- ggplot(freedom.care, aes(x=factor(group_name), y=quiz))

care.q <- care.q + geom_boxplot(aes(fill = group_name)) + theme_economist_white() + theme(legend.positi
  theme(axis.title.x=element_blank()) + labs(y = "Quiz Scores")

grid.arrange(care.c, care.q, ncol=2)

freedom.corp <- subset(freedom, freedom$channel=="Corp-retail")

corp.c <- ggplot(freedom.corp, aes(x=factor(group_name), y=comp_rate))

corp.c <- corp.c + geom_boxplot(aes(fill = group_name)) + theme_economist_white() + theme(legend.positi
  theme(axis.title.x=element_blank()) + labs(y = "Completion Rate") +
  #theme(axis.text.x = element_text(angle = 0, hjust = -1)) +
  scale_x_discrete(labels=c("Clgry", "Edmon", "Ottawa", "ON East", "ON Cen",
    "ON SE", "ON SW", "Van East", "Van West"))

corp.c
#grid.arrange(corp.c, corp.c, ncol=2)

#CORPCA01???Calgary CORPED01???Edmonton CORPON01???Ottawa CORPON02???ON East CORPON03???ON Central CORP

```

```

freedom.deal <- subset(freedom, freedom$channel=="Dealer-retail")

corp.d <- ggplot(freedom.deal, aes(x=factor(group_name), y=comp_rate))

#DEALCA01DEALEDO2DEALON01DEALON02DEALON03DEALON04DEALON05DEALON06DEALVA01

corp.d <- corp.d + geom_boxplot(aes(fill = group_name)) + theme_economist_white() + theme(legend.position = "top") +
  theme(axis.title.x=element_blank()) + labs(y = "Completion Rate") +
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5)) +
  scale_x_discrete(labels=c("CA01", "ED02", "ON01", "ON02", "ON03", "ON04", "ON05", "ON06", "VA01"))

corp.d
#grid.arrange(corp.c, corp.d, ncol=2)

#CORPCA01???Calgary CORPED01???Edmonton CORPON01???Ottawa CORPON02???ON East CORPON03???ON Central CORPON04???ON North

correl <- read.csv("correl.csv")

p <- qplot(as.Date(as.character(correl$Day), "%Y%m%d"), correl$Completion, data=correl, xlab="Day", ylab="Completion Rate", main="Completion Rate by Day")
p

meanChannel = aggregate(list(mean = freedom$comp_rate), list(channel = factor(freedom$channel)), mean)

STDChannel = aggregate(list(std = freedom$comp_rate), list(channel = factor(freedom$channel)), sd)

channel <- merge(meanChannel,STDChannel, by = "channel")

meanCare = aggregate(list(mean = freedom.care$comp_rate), list(group = factor(freedom.care$group_name)), mean)

STDCare = aggregate(list(std = freedom.care$comp_rate), list(group = factor(freedom.care$group_name)), sd)

Care <- merge(meanCare,STDCare, by = "group")

meanDeal = aggregate(list(mean = freedom.deal$comp_rate), list(group = factor(freedom.deal$group_name)), mean)

STDDeal = aggregate(list(std = freedom.deal$comp_rate), list(group = factor(freedom.deal$group_name)), sd)

Deal <- merge(meanDeal,STDDeal, by = "group")
#grid.arrange(channel, Care, Deal, ncol=2)

channel
Care
Deal

m1 <- lm(quiz ~ comp_rate, data = freedom)
m1

```

```

p <- qplot(freedom$quiz, freedom$comp_rate, data=freedom, xlab="Completion Rate", ylab="Quiz Score", ma

qqnorm(m1$residuals) + theme_economist_white()
qqline(m1$residuals) + theme_economist_white()
p

#grid.arrange(p, p, ncol=2)
#p + abline(lm(freedom$quiz ~ freedom$comp_rate))

#plot(freedom$quiz ~ freedom$comp_rate) + theme_economist_white()
#abline(m1)

m1 <- lm(Completion ~ as.Date(as.character(correl$Day), "%Y%m%d"), data = correl)
m1
summary(m1)

p <- qplot(as.Date(as.character(correl$Day), "%Y%m%d"), correl$Completion, data=correl, xlab="Day", ylab="Completion Rate")

qqnorm(m1$residuals) + theme_economist_white()
qqline(m1$residuals) + theme_economist_white()
p

aovTest1 <- aov(comp_rate ~ channel, data = freedom)
summary.aov(aovTest1)

aovTest2 <- aov(comp_rate ~ group_name, data = freedom.care)
summary.aov(aovTest2)

aovTest3 <- aov(comp_rate ~ group_name, data = freedom.corp)
summary.aov(aovTest3)

aovTest3b <- aov(comp_rate ~ group_name, data = freedom.deal)
summary.aov(aovTest3b)

m1 <- lm(quiz ~ comp_rate, data = freedom)
summary(m1)

posthocPairwiseT( aovTest1, p.adjust.method = "bonferroni")

posthocPairwiseT( aovTest2, p.adjust.method = "bonferroni")

posthocPairwiseT( aovTest3b, p.adjust.method = "bonferroni")

print("Care channel")
inference(y=freedom.care$comp_rate, est="mean", type="ci", method="theoretical", graphics.off())
print("Corporate Retail channel")
inference(y=freedom.corp$comp_rate, est="mean", type="ci", method="theoretical")
print("Dealer Retail channel")
inference(y=freedom.deal$comp_rate, est="mean", type="ci", method="theoretical")

```

```

temp = aggregate(list(comp_rate = freedom$comp_rate), list(channel = factor(freedom$channel)), mean)

mean1 <- ggplot(temp, aes(x=channel, y = comp_rate, fill = channel)) + geom_bar(stat = "identity") +
  theme_economist_white() +
  theme(axis.title.x=element_blank()) + theme(legend.position="none") + labs(y = "Completion Rate") +
  geom_text(aes(label=round(comp_rate,3))) +
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5)) +
  scale_x_discrete(labels=c("Care", "Corp", "Deal"))

temp2 = aggregate(list(comp_rate = freedom.care$comp_rate), list(group_name = factor(freedom.care$group_name)), mean)

mean2 <- ggplot(temp2, aes(x=group_name, y = comp_rate, fill = group_name)) + geom_bar(stat = "identity") +
  theme_economist_white() +
  theme(axis.title.x=element_blank()) + theme(legend.position="none") + labs(y = "Completion Rate") +
  geom_text(aes(label=round(comp_rate,3))) +
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5)) +
  scale_x_discrete(labels=c("Star", "Tele", "Wind"))

temp3 = aggregate(list(comp_rate = freedom.deal$comp_rate), list(group_name = factor(freedom.deal$group_name)), mean)

mean3 <- ggplot(temp3, aes(x=group_name, y = comp_rate, fill = group_name)) + geom_bar(stat = "identity") +
  theme_economist_white() +
  theme(axis.title.x=element_blank()) + theme(legend.position="none") + labs(y = "Completion Rate") +
  geom_text(aes(label=round(comp_rate,2))) +
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5)) +
  scale_x_discrete(labels=c("C1", "E2", "O1", "O2", "O3", "O4", "O5", "O6", "V1"))

grid.arrange(mean1, mean2, mean3, ncol=2)

##

```