STQD6134: Business Analytics

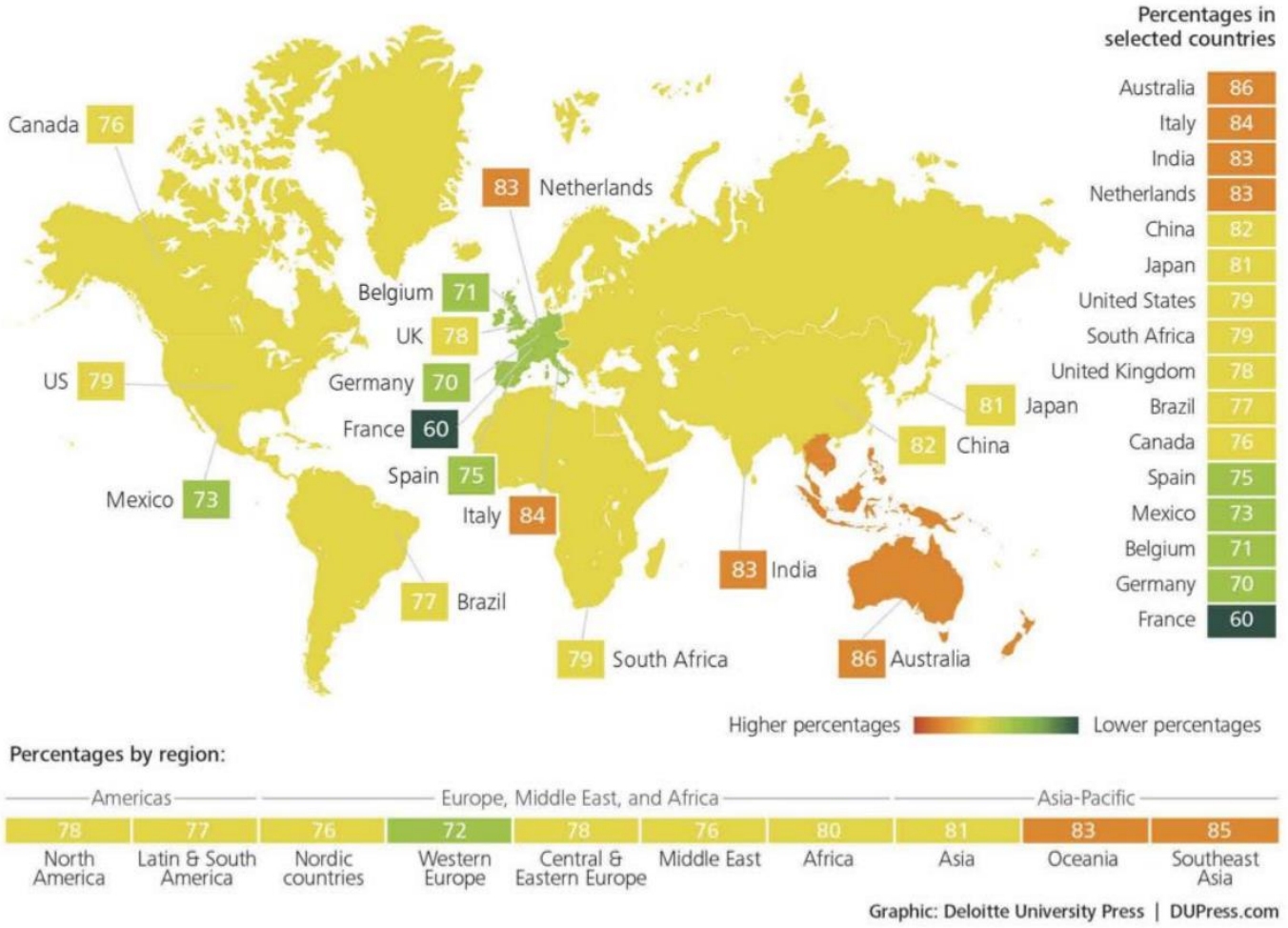
Time Series Analysis: Employees' Assessment & Forecasting

(Part I)

Human resource analytics

- In 2015, only 8% of the worldwide company are applying human resource analytics (Deloitte's Human Capotal Trend, 2016)

Figure 1. People analytics: Percentage of respondents rating this trend "important" or "very important"



Examples of human resource analytics

1. **Pay for performance**

There has been a steady and constant shift towards ensuring that what employees are paid is closely tied to their contribution to the organization. Some organizations are removing the artificial limits that kept high performers from earning more than their managers. For a number of the companies we work with the alignment of pay for performance is the number one shared agenda item between the CEO and the head of HR.

One of the best ways to demonstrate this practice is through a metric called the "Performance based compensation differential". This metric expresses how much more high performers are paid compared to their average performing peers. For example, a score of 1.2 means that on average high performers receive 20% more compensation than average performers. Turning this critical question into a single number allows for powerful insight across the organization; it means that different locations, business units and groups of employees can be easily compared using simple visual analyses.

One powerful way this translates to business value is during the annual pay review cycle. Most HRIS systems allow you to enter changes in pay, however these systems do not enable you to analyse how these awards relate to performance and whether or not they are aligned to the goals of the organization. Most HR departments provide guidance and then trust their managers to get it right.

Being able to analyse all of these decisions in real time, report this back to the organizational leadership and then revise these adjustments before they are confirmed leads to a demonstrated ability to ensure that the budget increases going into labour costs are being applied in the optimal way.

2. Prediction

It is one thing to know what has happened in HR - the majority of HR data to date has focused on the reporting of transactional outcomes - but another thing to know what *will* happen. For example, lots of HR groups report the percentage of people that had a performance review or who completed an engagement survey. This type of reporting relates to the process orientation of HR and, although interesting, does little to demonstrate the true value of the functions.

In addition, the opportunities to add value through HR practices come more from stopping the wrong outcome from happening, than from reporting on what has happened. For example, the cost of voluntary turnover has been established at approximately 1.5 times annual base pay for salaried employees (Source: PWC Saratoga and CEB). Therefore, if you prevent two high value employees, with salaries of 50,000GBP, from leaving the organization you have saved approximately 150,000GBP. In order to achieve this type of saving you need to know who will leave, before they have left.

This is where sophisticated algorithms, that use historical data to determine the likelihood that someone will resign, come into play. There are a number of known actions that will prevent someone from leaving like signing bonuses, formal agreements around career progression and learning opportunities. However, the crucial part is knowing to whom you should offer these incentives. When it is possible to focus on the right population, through powerful and validated statistical models, it leads to better outcomes for lower cost.

Another place where prediction is becoming valuable and important for the companies we work with is in relation to retirement. The pattern of behavior relating to retirement is changing with more and more people delaying retirement or shifting to contract or part-time work than ever before. Prediction here is important as often the people retiring are in key

roles or hold key relationships and are critical for the business to ensure continuity of performance. However, it is also challenging to keep a potential successor waiting if the incumbent chooses not to retire at the time expected.

Instead of using the old indicators of age and tenure to estimate retirement behavior modern analytics technology applies algorithms that take into account many additional factors such as recent changes in role, pay level, rates of change in pay and incentive eligibility to refine the prediction of who will retire. This allows companies using this type of analytic approach to be more successful and effective in managing the retirement cycle and ensuring that key roles have a successor ready at the right time.

3. Retention

It is common for retention issues to have a stop/ start response where money is thrown at employees in the hope that it will change the outcome. More often than not, this approach does not work or has only a very limited and short-term effect. The money spent here does not deliver the outcomes expected.

Our customers report improved retention outcomes, as well as a better ability to focus resources and programs where they need them. They achieve this through the use of common metrics such as turnover, resignation, involuntary turnover, etc. However, the differentiator is the ability to compare trends over time, across business units or between key groups of employees to the overall organizational outcomes. It is not the standalone metrics that brings the insight, but the ability to quickly build comparisons, identify trends and find outliers that makes the difference.

In addition, these companies are using our clustering algorithms to determine the common features of employees that are related to higher or lower retention rates. This insight means the right approach can be taken with the right employees, leading to better results at a lower overall cost.

Real life examples

1. Google

In his book Work Rules! (2015), Laszlo Bock, Senior Vice President of People Operations (HRM) at Google, writes that the most important instrument of Google's People Operations is statistics. The questions interviewees get asked in Google's hiring process are all fully automated, computer-generated and fine-tuned in order to find the best candidate. On top of that, Google estimates the probability of people leaving the company by applying HR predictive analysis. One of Google's findings is that new salespeople, who do not get a promotion within four years, are much more likely leave the company.

2. Facebook pages

Do your recruiters check the Facebook pages of applicants? Maybe they should. A 2012 study revealed that it is possible to predict someone's personality and future work performance based on their Facebook profile (Kluemper, Rosen & Mossholder, 2012). In this study, a number of participants gave hirability ratings based on Facebook profiles. These ratings predicted 8% of manager-rated job performance for these people.

8% is not that much. For instance, a standard personality test has a higher predictive value for performance compared to looking at someone's Facebook profile. However, the literature shows time and time again that the best predictive models for future job performance combine various predictors, such as IQ tests, structured interviews, and personality tests together. Looking through a Facebook profile could be an additional instrument to scan candidates.

3. US Special Forces

During the highly selective training, the U.S. Special Forces predict which candidates are most likely to succeed. Two key predictors are ‘grit’ and the ability to do more than 80 pushups. Grit was actually a more accurate predictor of training success than IQ. Check Angela Lee Duckworth’s Ted Talk if you’re interested to know more.

4. Wikipedia

Wikipedia editors, or Wikipedians, create and edit articles to keep the world’s largest encyclopedia up-to-date. Each day, over 800 new pages are created and 3,000 edits are made on the English Wikipedia alone. Wikipedia is able to predict who of its 750,000 editors is most likely to stop contributing.

5. Best Buy

Best Buy (a leader in HR predictive analytics) can accurately predict how employee engagement impacts the performance of their stores. A 0.1% increase in employee engagement results in an increase of over $100,000 in the store's annual income. The enormous impact of engagement prompted Best Buy to make its engagement surveys quarterly instead of annually. Measuring the impact of employee engagement on bottom line performance is difficult to do but certainly possible, as this example shows.

6. Never hire toxic people

This is a case-study published by Cornerstone (2015). Cornerstone studied the impact of toxic employees on the workplace. Toxic employees are employees who are most likely to engage in toxic behavior. Examples of these behaviors are fraud, drugs or alcohol abuse, and sexual harassment.

These people are not only damaging to the company; they are highly toxic to the general work environment. Previous research suggested that one toxic employee in a team would cause productivity to decrease by 30% to 40%. On top of that, good employees are more likely to quit when they have to work together with toxic colleagues.

Cornerstone used a dataset of 63,000 employees. In this dataset, they marked which employees were involuntarily terminated due to workplace violence, falsification of documents, drugs, and alcohol abuse, and other policy violations. Based on these criteria, around 4% of all employees could be classified as being 'toxic'.

After analyzing the dataset, Cornerstone identified a number of key characteristics of toxic people.

Toxic people:
1. are self-proclaimed rule-followers;
2. score low on attendance and dependability;
3. and have a low service orientation.

Remarkably, the study did not find the previously reported high levels of productivity loss in the short term. However, it did find toxic behavior to be contagious. People who work together with toxic colleagues are also more likely to quit. Additionally, the study hypothesized that toxic colleagues contribute to long-term stress and burnout among other employees.

In the end, Cornerstone proved that hiring a toxic employee will cost the employer $12,800 on average, versus an average of $4,000 for a non-toxic employee. This excludes the long-term (and costly) productivity loss through burnout and other negative effects. By fine-tuning the hiring process, companies can prevent hiring candidates who are likely to become toxic and create a healthier working environment.

Thinking exercise

1) If you are given a set of data of a company's employees profile, and you want to identify the best and worst employees for an award (and to be sacked); what is your plan on analyzing the data?

2) After consulting with their HRIS staff, they found that they have access to the following information:

- EmployeeID
- Record Date
- Birth Date
- Original Hire Date
- Termination Date (if terminated)
- Age
- Length of Service
- City
- Department
- Job title
- Store Name
- Gender
- termination reason
- termination type (voluntary or involuntary)

Case study: Forecasting Future Ridership

The finance group approached the BI team and asked for help with forecasting future trends. They heard about your great work for the marketing team and wanted to get your perspective on their problem.

Once a year they prepare an annual report that includes ridership details. They are hoping to include not only last year's ridership levels, but also a forecast of ridership levels in the coming year. These types of time-based predictions are **forecasts**. The `Ch6_ridership_data_2011-2012.csv` data file is available at the book's website at `http://jgendron.github.io/com.packtpub.intro.r.bi/`.

This data is a subset of the bike sharing data you used in the first two chapters. It contains two years of observations, including the date and a count of users by hour. Your task is to convert this data into a time series aggregated by month, apply a time series model, and produce a monthly forecast for the coming year.

Analyzing time series data with linear regression

Before working with the data from the use case, we will use a dataset already in R. The `TSA` package contains a dataset called **airpass**. This dataset provides the total monthly count of international airline passengers covering the period from January 1960 to December 1971. This represents twelve years of monthly passenger data, which is 144 observations. After loading the library, the `airpass` dataset is available using the `data()` function. You can examine the dataset using methods discussed in the previous chapters:

```
library(TSA)
data(airpass)
str(airpass)
summary(airpass)
```

The output is as follows:

```
Time-Series [1:144] from 1960 to 1972: 112 118 132 129 121 135 ...
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  104.0   180.0   265.5   280.3   360.5   622.0
```

This is a time series. This means that responses are dependent on previous points in time. *This dependent data would fail the assumption of independence*. What would a linear regression look like anyway? The following is code to create a temporary data frame to generate a linear model. Begin by extracting passenger volume and assigning the single column as a matrix to the `volume` variable:

```
volume <- as.matrix(airpass)
```

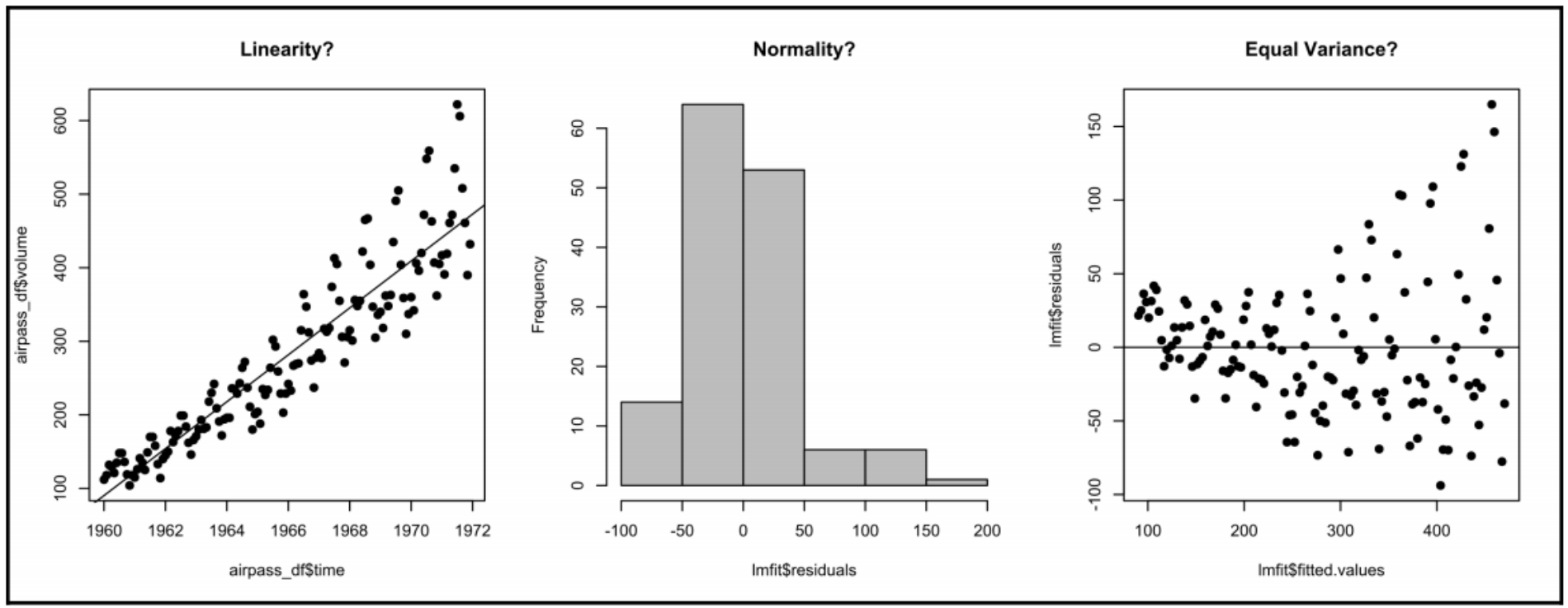
```
time <- as.matrix(time(airpass))
```

```
airpass_df <- as.data.frame(cbind(volume, time))
colnames(airpass_df) <- c("volume", "time")
```

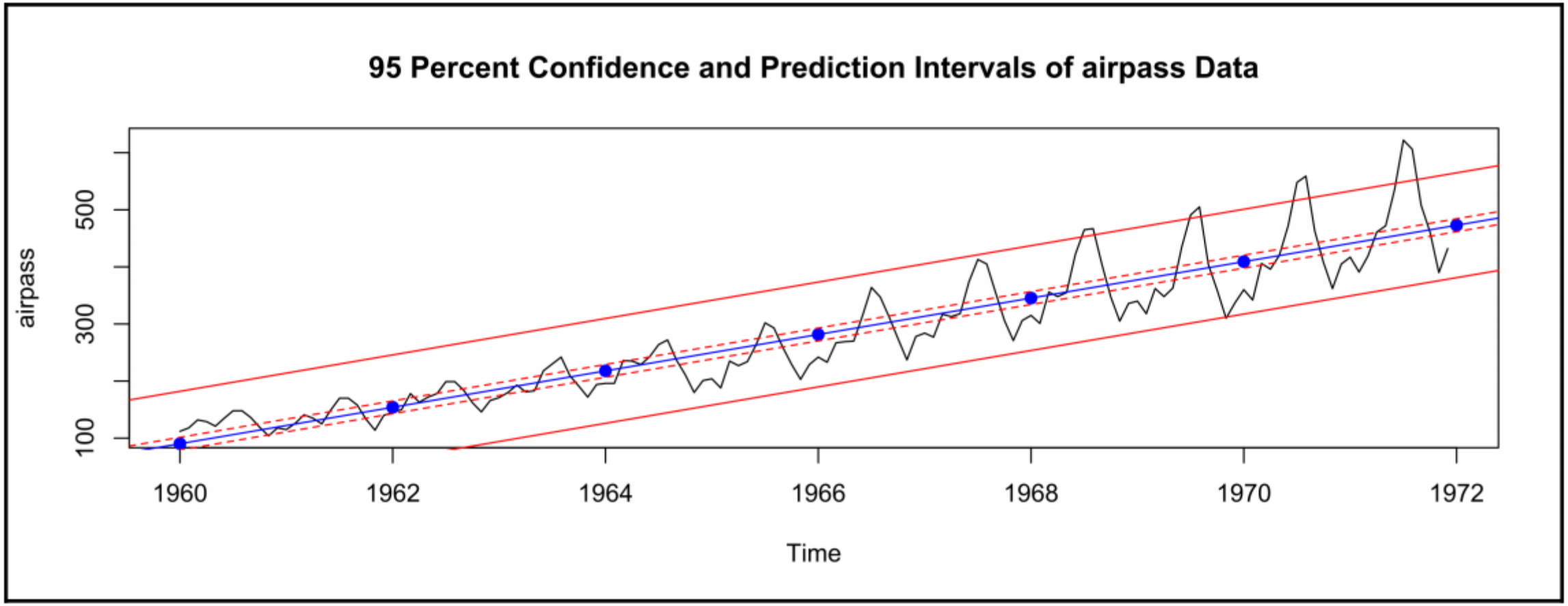
```
lmfit <- lm(volume ~ time, data = airpass_df)
summary(lmfit)
```

```
par(mfrow = c(1, 3))
plot(airpass_df$time, airpass_df$volume, pch = 19, main = "Linearity?")
abline(lmfit)
hist(lmfit$residuals, main = "Normality?", col = "gray")
plot(lmfit$fitted.values, lmfit$residuals, main = "Equal Variance?",
     pch = 19)
abline(h = 0)
```

The output is shown as follows:



```
plot(airpass, main = "95 Percent Confidence and Prediction Intervals of
airpass Data")
abline(lmfit, col = "blue")
newdata <- data.frame(time = seq(1960, 1972, 2))
pred <- predict.lm(lmfit, newdata, interval = "predict")
points(seq(1960, 1972, 2), pred[ ,1], pch = 19, col = "blue")
abline(lsfit(seq(1960, 1972, 2), pred[ ,2]), col = "red")
abline(lsfit(seq(1960, 1972, 2), pred[ ,3]), col = "red")
pred <- predict.lm(lmfit, newdata, interval = "confidence")
abline(lsfit(seq(1960, 1972, 2), pred[,2]), lty = 2, col = "red")
abline(lsfit(seq(1960, 1972, 2), pred[,3]), lty = 2, col = "red")
```



The linear model is a terrible predictor of these dependent, time-based values. All predictions will fall on the center line-the trend. It will not capture the predictable cycles that you see. In addition, some forecasts systemically appear outside the confidence interval, which indicates an underlying issue with the model. You will need to use another type of regression called time series analysis.

Forecasts versus predictions: In linear regression, you made **predictions**. You predicted a response using other variables-not previous responses. With time series analysis, you will make a forecast. The **forecast** is a number of future responses, based on the previous responses themselves. Consider the price of a house. Here is how you would make a prediction and forecast:

- **Prediction**: You can predict the sales price based on variables that influence the response, such as crime, access to transportation, and schools
- **Forecast**: You can forecast the sales price based on a time series model using the previous sales prices themselves to forecast future prices