HR Predictive Data Analytics Requirements Specification

Version 3.0

**Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Date** | **Reason For Changes** | **Version** |
| Initial Draft | 09/20/2017 | Unavailability of the Test file is needed for each task, including test actions, objectives, and expected results | 1.0 |
| Second Draft | 11/26/2017 | Change in the product requirements by the customer | 2.0 |
| Final Draft | 12/06/2017 | Final changes to match the exact requirements of the user | 3.0 |

**Document Sign Off**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Date** | **Signature** | **Role** |
| **Celine Dimbo** | **12-10-2017** | **Celine Dimbo** | Analyst |
| **Saketh Gourisetty** | **12-10-2017** | **Saketh Gourisetty** | Developer |
| **Shama Zabeen Shaik** | **12-10-2017** | **Shama Zabeen Shaik** | Developer |
| **Venkatesh Uma Maheswaran** | **12-10-2017** | **Venkatesh Uma Maheswaran** | Developer |
| **Vinoth SivaKumar** | **12-10-2017** | **Vinoth SivaKumar** | Analyst |
| **Yue Peng** | **12-10-2017** | **Yue Peng** | Manager |

**Table of Contents**

[**1.**](#_2s8eyo1) **Executive Summary 4**

[1.1](#_17dp8vu) Project Overview 4

[1.2](#_3rdcrjn) Theoretical backgrounD 4

[1.3](#_3rdcrjn) learning experience 4

[**2.**](#_26in1rg) **Product/Service Description 5**

[2.1](#_lnxbz9) scope 5

[2.2](#_35nkun2) Assumptions 6

[2.3](#_1ksv4uv) Constraints 6

[2.4](#_44sinio) Agile methodoloGy of project development 6

[**3.**](#_2jxsxqh) **Methods 6**

[3.1](#_z337ya) sprint schedule 7

[3.2](#_3j2qqm3) User diagram 7

[3.3](#_1y810tw) context diagram 8

[3.4](#_4i7ojhp) data flow diagram level 1 diagram 9

[**4.**](#_2xcytpi) [**REQUIREMENTS**](https://docs.google.com/document/d/1EtvFyPhssnPe4HbYRSBsHLeWa2iIvtUtv2seL-RgxWw/edit#heading=h.3rdcrjn) 10

4[.1 Functional Requirements](https://docs.google.com/document/d/1EtvFyPhssnPe4HbYRSBsHLeWa2iIvtUtv2seL-RgxWw/edit#heading=h.28h4qwu) 11

4[.2 User Interface Requirements](https://docs.google.com/document/d/1EtvFyPhssnPe4HbYRSBsHLeWa2iIvtUtv2seL-RgxWw/edit#heading=h.lnxbz9) 11

4[.3 Usability](https://docs.google.com/document/d/1EtvFyPhssnPe4HbYRSBsHLeWa2iIvtUtv2seL-RgxWw/edit#heading=h.35nkun2) 11

4[.4 Performance](https://docs.google.com/document/d/1EtvFyPhssnPe4HbYRSBsHLeWa2iIvtUtv2seL-RgxWw/edit#heading=h.44sinio) 12

4[.4.1 Capacity](https://docs.google.com/document/d/1EtvFyPhssnPe4HbYRSBsHLeWa2iIvtUtv2seL-RgxWw/edit#heading=h.2jxsxqh) 12

4[.4.2 Availability](https://docs.google.com/document/d/1EtvFyPhssnPe4HbYRSBsHLeWa2iIvtUtv2seL-RgxWw/edit#heading=h.z337ya) 12

4[.4.3 Latency](https://docs.google.com/document/d/1EtvFyPhssnPe4HbYRSBsHLeWa2iIvtUtv2seL-RgxWw/edit#heading=h.3j2qqm3)

5. User Stories

6.  [Deleted or Deferred Requirements](#_Toc191724261)

7. Burndown Chart

8. **results 13**

5[.1](#_z337ya) sprint schedule 14

5[.2](#_3j2qqm3) User diagram 15

5[.3](#_1y810tw) context diagram 15

5[.4](#_4i7ojhp) data flow diagram level 1 diagram 16

[**ANNEXURE A. List of figures**](#_1ci93xb)

[figure 1.1 sprint schedules 11](#_1y810tw)

[figure 1.2 user diagram 12](#_4i7ojhp)

[figure 1.1 context diagram 12](#_z337ya)

[figure 1.3 Data flow diagram level 1 diagram 13](#_3j2qqm3)

[figure 1.1 age calculation formula 13](#_3j2qqm3)

[figure 1.2 age calculation results 14](#_1y810tw)

[figure 1.3 importing excel data into tableau 14](#_4i7ojhp)

[figure 1.4 excel data selection into tableau 15](#_z337ya)

[figure 1.5 excel file selection 15](#_3j2qqm3)

[figure 1.6 new worksheet selection in tabelau 16](#_1y810tw)

[figure 1.7 tableau bar chart 17](#_4i7ojhp)

[figure 1.8 frequency selection in tableaue 18](#_z337ya)

[figure 1.9 tableau line graph 19](#_3j2qqm3)

[figure 1.10 tableau forecast 20](#_1y810tw)

[figure 1.11 show forecast selection 21](#_4i7ojhp)

[figure 1.12 forecast selection 22](#_3j2qqm3)

[figure 1.13 model and time period selection 23](#_1y810tw)

[figure 1.14 trend line selection in tableau 24](#_1y810tw)

[figure 1.15 crosstab to excel 25](#_4i7ojhp)

[figure 1.16 data exported into excel with labels 26](#_z337ya)

[figure 1.17 data loading in r 27](#_3j2qqm3)

[figure 1.18 percentage of employees leaving the company 28](#_1y810tw)

[figure 1.19 business unite based attrition 29](#_4i7ojhp)

[figure 1.20 termination types and status year 30](#_3j2qqm3)

[figure 1.21 termination by status year and reason 31](#_1y810tw)

Fi[gure 1.22 termination by reason and department 32](#_3j2qqm3)

[figure 1.23 termination by age and department 33](#_1y810tw)

[figure 1.24 age anage length of service distributrion by status 34](#_4i7ojhp)

[figure 1.25 partitioning the dataset 34](#_1y810tw)

[figure 1.26 decition tree 35](#_4i7ojhp)

[figure 1.27 data training 35](#_1y810tw)

[figure 1.28 pie chart for the factors 36](#_4i7ojhp)

[figure 1.29 testing the trend data based on status and age 36](#_1y810tw)

[figure 1.30 based on length and service 36](#_4i7ojhp)

[figure 1.31 based om status and gender 37](#_1y810tw)

[figure 1.32 based on status and year 37](#_4i7ojhp)

[figure 1.25 partitioning the dataset 37](#_1y810tw)

[figure 1.26 decition tree 36](#_4i7ojhp)

[figure 1.27 data training 36](#_1y810tw)

[figure 1.28 pie chart for the factors 36](#_4i7ojhp)

[figure 1.29 testing the trend data based on status and age 37](#_1y810tw)

[figure 1.30 based on length and service 37](#_4i7ojhp)

[figure 1.31 based om status and gender 37](#_1y810tw)

[figure 1.32 based on status and year 38](#_4i7ojhp)

[figure 1.33 based on status and location 38](#_1y810tw)

[figure 1.34 based on status age and gender 38](#_4i7ojhp)

**1. EXECUTIVE SUMMARY**

* 1. **Project Overview**

HR predictive modeling project was done using data provided by Continental AG. The project involves creating Dashboard for the executives and supervisors to view predictive models and analytics data. Continental is in the process of evaluating various tools, software, and techniques for creating dashboards with reporting and predictive models. There was no restriction on the usage of any software, tool, or algorithm to come up with predictive models on the company data. We also had to provide the data points we would require making these predictions. The projects major tasks were identifying data, cleaning data, generating predictive models, displaying those models on a dashboard along with identifying the various tools which could do these above tasks. The project addressed these challenges coming up with dashboard by carrying out all the above activities.

Our Team worked with Dr. Mohsen Dorodchi, Professor, College of computing and informatics UNCC, Mr. Dave Devore and Mr. Mytrik Shah from Continental the customer and stakeholder who would also provide data from continental and review our dashboard to provide feedback on its alignment to their requirements. In the course “Software system design and implementation”. To design this dashboard which provides graphs and data of predictive models on attrition rates, time to hire, and leave of absence data sets. We also made sure these dashboards have various permission criteria based on which information segregation is done to show only specific data for which the user has permission.

**1.2 Theoretical Background**

The major business concerns for every organization nowadays are Employee attrition and candidate absconding. Employee attrition is a very serious issue, because employees are the most important human capital assets in today’s marketplace. Attrition impacts the competitive advantage of an organization. Due to employee attrition there will be may additional costs that impacts the productivity of the organization, they are: Cost of training new employees, the recruitment and selection costs, adjustment time, possible product and/or service quality problems, costs of temporary staff, the cost of training, the cost for loss of productivity, the cost of lost knowledge and the cost of the position remaining vacant till a suitable replacement is found.

The profitability and the competitive advantage of the firm are significantly taken back away with the costs. It is very essential for every organization to understand what factors restore a potential candidate and predict the attrition in the early stages of the recruitment cycle, in order to ensure a non-significant loss of the productivity among the recruiters, the hiring managers and also loss of revenue and money. Employee attrition is a global level problem. Due to the constant increase in the attrition rate these days, the software industry is been effected the most. The continued reports of the attrition rate in a company, the effects of the attrition on the revenue and the growth, and the requirement of man force in the recording of the number of people leaving the company, has interested us in coming up with a model that makes use of the existing data records to train itself and test for the future attrition value.

## **1.3 Learning Experience**

The goal of this project was to come up with a solution evaluating the various components from multiple vendors available and integrating the best fit methods and tools to create the final product. Our team was a mix of individuals from various parts of the world. The working experience was at par with industry experience working with multinational team. We had to do this project using the Agile framework which was new to all of us. The framework was helpful in the process as we were to evaluate different tools and iteratively develop the project. So agile gave us the flexibility to do changes on the go based on our experience and decide the mix of tools to be used for the final product. We got to learn about R, SAS and Tableau for data cleaning and analytics, and Plotly, Kibana for the dashboard. The project needed project management, data analytics, visualization and programming skills. This project provided a platform to apply knowledge from the course work in Information technology project management to manage the project in terms of scope, time and effort, to have milestones within projects, assign roles and responsibilities, look for risks and mitigate the same. Business intelligence and analytics and Knowledge discovery in database knowledge helped to understand what data is required for models, methods and process to clean data, and come up with predictive models. Principles of human computer interaction was important to know how to design a dashboard starting with wireframe designs and low fidelity prototypes moving on to the final product which follows the principles of design and create an efficient and easy to use dashboard.

# 2. Product/Service Description

Continental as a German automotive manufacturing company specializing in manufacturing of tires, break systems, powertrain, chassis and other parts for the automotive transport industry. The company wanted to work with Graduate level students and come up with dashboards to visualize their company data. Our team was involved in creating dashboard using predictive data. The data source was HR data provided by the company. This initiative was very important for Continental as they were relatively new to the process of creating such dashboards. The goal was to leverage the power of data analytics to crunch large amount of unstructured data and come up with meaningful insights in the fields of attrition, Time to hire and absenteeism. This would not only provide ability to see what happed but also to look ahead into what is about to happen. Such information would be used by the top executives for better business decisions. Ours was first of this kind analysis done on continental HR data so it was important for us to create a good foundation highlighting the ability and usefulness of information gained from their data on which further improvements can be done. The team was provided the freedom to choose tools for predictive analysis, data cleaning and visualization. This gave us the opportunity to try a variety of tools before finalizing on the products to use.

## **2.1 Scope**

Following were the scope of the project.

* Use Agile methodology to develop the project. Come up with project plan, sprint plan, allocate member roles.
* Create database to import data provided. Apply data cleaning methodologies to preprocess the data.
* Shortlist data analytics tools to be used. Come up with algorithms to process the data related to attrition rate, time to recruit, absenteeism.
* Shortlist visualization tools for the dashboard. Create dashboard with predictive data

## **2.2 Assumptions**

* The HR Data to be provided by continental. The data would be related to attrition, absenteeism, and time to recruit.
* The data provided would have quantity and quality required to do predictive analytics.
* Data owner would help providing explanation of attributes in the data set.
* Continental does not have any restrictions on the tools to be used for the project.
* Only two levels of users need access to the dashboard. Executive and managers.

## **2.3 Constraints.**

* The project must be completed during the Fall 2017 semester from august to December 2017
* Teams size fixed at 6 members.
* The members did not have dashboard creation expertise.
* Agile methodology had to be used for project development.

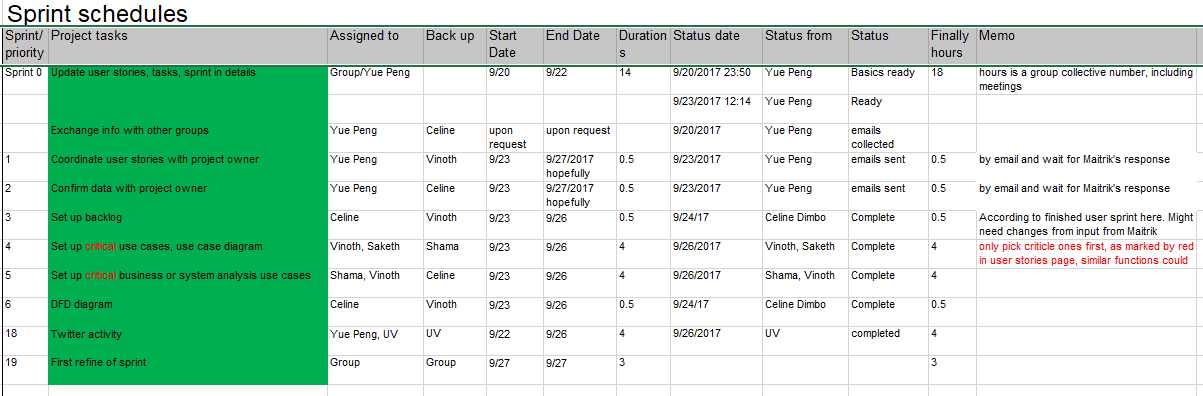
## **2.4 Agile methodology of project development**

The traditional methodology the waterfall development method is an approach by which the different stages are completed one after another. specific outputs from previous stage act as inputs to the next stage. Like 1) Requirements, 2) Design, 3) Development, 4) Testing, 5) Maintenance. Requirement analysis is completed and that is used as inputs to the Design stage. Once design is completed the development will use the inputs for develop the product which is then tested and passed on to maintenance. Agile manifesto is different from waterfall approach. It gives priority to individuals and interactions over processes and tools, working software over comprehensive documentation, customer collaboration over contract negotiation, responding to change over following a plan. The project is split into sprints and each sprint has specific tasks. All the above tasks mentioned can be done simultaneously. One can always revisit design even during development and design phase. And parts of the projects are completed in each sprint. The philosophy is instead of large group spending long time building big things a small team spending short time to build small things and integrating it regularly.

# 3. Methods

The course work of SSDI had lots of lots of new methods and techniques which would be taught both theoretically which was put to practical use. The information technology project management (ITCS6112) course was useful to understanding the scope, time, cost constraints of this project. It helped in creating work breakdown structure of the various tasks associated with the project. The project had 3 main components with multiple smaller tasks. The project management involved creation of teams for each task assign responsible and backup for each task. With a team of 6 individuals from different cultural background and working style the experience was as good as a real multinational industry project. Time management was very important as the project was of limited period. Each member had various classes and other deliverables which made it a challenge to get things done. The time management techniques thought in the coursework were used in this project to stick to the time lines for completion. Following is a sprint sheet with information on tasks and other attributes used in the project.

Figure 3.1 Sprint Schedules

****

The course work provided opportunities to work with data analytics tool like R, Weka, SAS. And database like MySQL. These courses were a doorway databases and data analytics along with the numerous techniques to be used for working with data. I got to learn about linear regression, logistic regression, classification, decision tree, clustering etc. The major task in our project was cleaning data and data transformation. There were lot of null and duplicate and unimportant fields which had to be eliminated. Statistical methods like range and clustering were used on different columns of data. Data transformation included getting state names from complex location in the data set, calculating total employment period, Age of employee, were also done from the raw data provided. Stats of few of the transformation done on the dataset are below

Remove Null Birth date (38%).

Remove unknown gender.

Remove functional area column (95% null)

Remove null cost center (0.5% data removed)

Remove unknown gender

Remove functional area column (96% null)

Remove null cost center (2% data removed)

Remove rating column (all null)

Remove PS group column (118 differences, 30.3% null)

Remove Position (99% Integration: default)

Remove PA and P Subarea (code for personal area)

Remove business unit and Division Code (57.63% null, 60 different values)

Remove Employee status (all withdrawn)

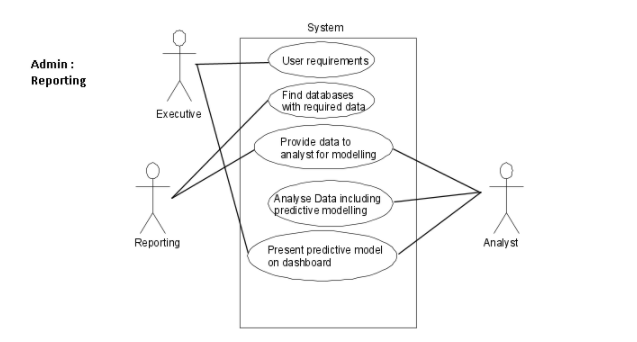
Remove Ethnicity column (50% null)

Remove duplicates (by manually deleting 60003821 60046184 60169779 60169951 60180289 60245267)

Personnel Area Split to country, states, cities, and subs

Remove sub area because it's the same with split area

US-ALL can be found in Personnel Subarea.

Figure 3.2 User Diagram****

Software system design and implementation (ITIS 6112). The project was done as part of coursework of software system design and implementation. The learnings about agile methodology of project implementation was applied in this project implementation. The project is based on the Model – View – Controller software architectural pattern thought in the network based application development course. The usage of GitHub thought in coursework was important in creating a repository for code versioning for seamless changes to codes with a chance to rollback if there were issues. The usage of context diagram, DFD diagram for project design and implementation was utilized for documentation purposes.

Figure 3.3 Context Diagram

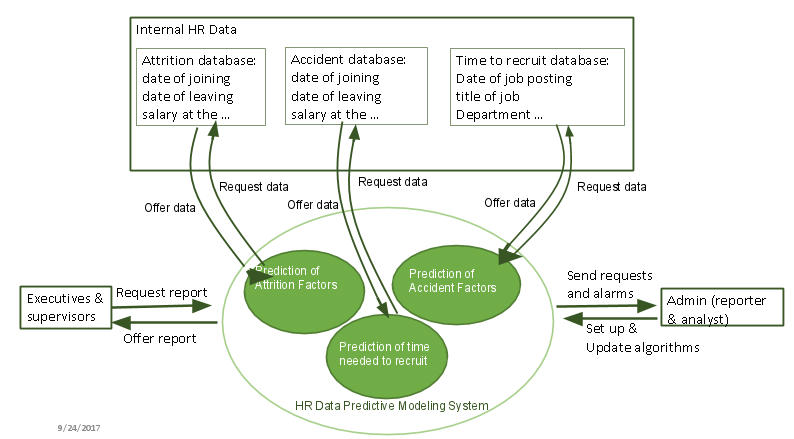
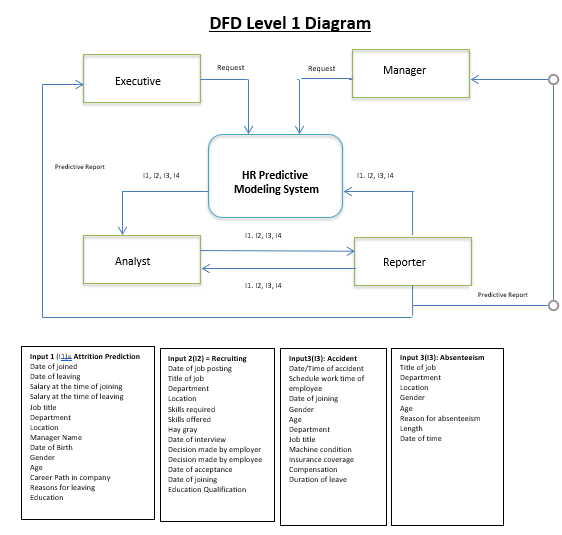


Figure 3.4 Data Flow Diagram Level 1 Diagram



**4.2 Requirement Analysis**

There are 2 types of system requirements, Functional Requirements and Non- Functional Requirements.

**4.2.1 Functional Requirements:**

The product is a dashboard reporting on specific predictive analytics inquiries regarding internal HR data. The objective is to explore the large cloud of HR data, discover information, provide meaningful predictions, and visualize data with clarity

**4.2.1.1 Product Perspective**

**4.2.1.2 Product Features**

* Switch between tabs to see different datasets
* Choose a specific time range
* All or a specific group of employees
* Choose how long to forecast
* Choose time base of forecasting
* Choose a specific age range
* Choose a specific gender
* All locations/no of data

**4.2.1.3 User characteristics**

Users are provided with a whole range of data to select from. From which, the selection enables the user to filter out as per the system requirements. And would be able to predict as per their requirements on the attrition, absenteeism, and time to hire rates.

**4.2.1.4 Assumption and Dependencies**

The assumptions that we have made are, the data provided by the company for the prediction is correct and accurate. Eliminating data would not affect in much of the transformations in the final predicted output. The data trained and tested is done right.

**4.2.1.5 User Requirements**

The user wanted different modules where in the Executive and the manager would be able to view the data on the modules, with the executive having the maximum viewability of the data and the manager has comparatively less data visibility. The analyst and the reporter are the people who have access to the code and the data manipulation methods.

**4.2.2 Non Functional Requirements**

**4.2.2.1 Product requirements**

* Use cases
* Data for the fields to predict
* Product outlook
* Project acceptance by the user
* Dashboard compatible software

**4.2.2.1.1 Efficiency**

Response time - The response time will be less than 12 seconds for 95% situational requests made to the system.

**4.2.2.1.2 Reliability**

Mean Time among Failures - The mean time among failures for the system will be 1000 hours.

**4.2.2.1.3 Portability**

Being available over the web(Having been launched on the server globally), could be easily ported to any location and could be very well used in the presence of internet connectivity.

**4.2.2.1.4 Usability**

The end users will be able to able to learn to use the system with a minimum training, as manual will be provided and it's just basic messaging.

**4.2.2.3 Engineering Standard Requirements**

· **Economic**

Open source softwares being used that are free for all the users. Not much investment needs to be made in the development and could be used without any specific requirements under just basic internet connectivity locations.

· **Environmental**

It has no degrading or negative effect on environment, also no enhancing effect also. It has a positive effect on it where we can see an increase in productivity of the company in all the significant regions.

· **Social**

It will help the users to use it even when travelling thus saving them a lot of time and helping them in socially improve their business and find the faults in their model to correct them for the organization’s well being.

· **Ethical**

Being an anti-theft software, it will help build the ethical base in society and instill a confusion among hackers.

· **Sustainability**

As mentioned above, it has a positive effect on environment; similarly it has a positive or no negative effect on sustainability.

· **Legality**

It can for sure help in ensuring effective implementation of law and order, and flaws in this system may be caught and put under law.

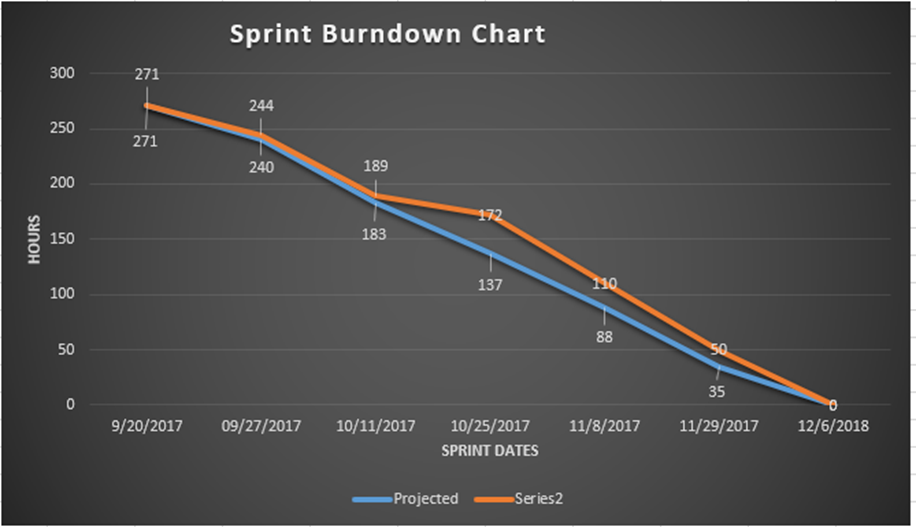
**5. User Stories:**

|  |  |  |
| --- | --- | --- |
| Index | Objectives | Data needed |
|  |  |  |
| 1 | see who are leaving and why | attrition data |
| 2 | see how long it takes to recuruit and the factors influencing | recruiting data |
| 3 | see who are absent and why | absence data |
| 4 | see the trends and reasons for accidents | accident data |
| 5 | get to the system functions |  |
| 6 | come back to my customizations |  |
| 7 | check the reports in my customized way and use my authorized functions | user data |
| 8 | customize my account | system data/user data |
| 9 | learn about how to use the system and display the data I want | system data |
| 10 | provide new requirements and ask for a feasibility study. | system data |
| 11 | check the situation of different historical time period or predictions for different short terms and long terms | result data |
| 12 | get alarmed if it meets a threshold or any problem emerges | system data |

6.  [Deleted or Deferred Requirements](#_Toc191724261)

|  |  |  |
| --- | --- | --- |
| **ID** | **Role** | **Story** |
| 1 | Executive and Manager | As a Executive/Supervisor I want to pull historical report/prediction on attrition/turnover of the company or a department so that I can see who are leaving and why |
| 2 | Executive and Manager | As a Executive/Supervisor I want to pull historical report/prediction on recruiting so that I can see how long it takes to recuruit and the factors influencing |
| 3 | Executive and Manager | As a Executive/Supervisor I want to pull History and predictions of absenteesim of the company or a department so that I can see who are absent and why |
| 4 | Executive and Manager | As a Executive/Supervisor I want to pull History and predictions of accidents of the company or a department so that I can see the trends and reasons for accidents |
| 5 | Executive and Manager | As a Executive/Supervisor I want to check/search help functions of the system for functions, databases, and reports so that I can learn about how to use the system and display the data I want |
| 6 | Executive and Manager | As a Executive/Supervisor I want to view/pull request of report of function on a discussion forum or ticketing system so that I can provide new requirements and ask for a feasibility study. |
| 7 | Executive and Manager | As a Executive/Supervisor I want to zoom/view on the timeline/range of any report in different scales so that I can check the situation of different historical time period or predictions for different short terms and long terms |
| 8 | Executive and Manager | As a Executive/Supervisor I want to set up alarm or warnings for updates of essential report so that I can get alarmed if it meets a threshold or any problem emerges |
| 9 | Reporter | As a Reporter, I want to register for the system and manage my profiles so that I can get to the system functions for me |
| 10 | Reporter | As a Reporter, I want to save my set ups in the system so that I can come back to my customizations |
| 11 | Reporter | As a Reporter, I want to login back in my account so that I can use the system in my customized way and use my authorized functions |
| 12 | Reporter | As a Reporter, I want to change customize my account and save them, including password, profiles, displays so that I can customize my account |
| 13 | Reporter | As a Reporter, I want to set up new users in the system, including accessibility of data, reports or functions of the system so that I can set correct access to new users |
| 14 | Reporter | As a Reporter, I want to check access/change the user database so that I can change the settings of different uses, such as access |
| 15 | Reporter | As a Reporter, I want to reply check the requests of the executive and supervisor in the forum, as well as requrests of data from the analyst so that I can provide the help they need |
| 16 | Reporter | As a Reporter, I want to acquire the requests of the executive and supervisor in the forum, as well as requrests of data from the analyst so that I can provide the help they need |
| 17 | Reporter | As a Reporter, I want to acquire data from primary and secondary data sources so that I can provide data to analyst |
| 18 | Reporter | As a Reporter, I want to change data to proper formats so that I can provide data with correct format to the analyst |
| 19 | Reporter | As a Reporter, I want to check updates of primary and secondary data sources so that I can update the database and reports |
| 20 | Analyst | As an Analyst, I want to register for the system and manage my profiles |
| 21 | Analyst | As an Analyst, I want to save my set ups in the system |
| 22 | Analyst | As an Analyst, I want to login back in my account |
| 23 | Analyst | As an Analyst, I want to change customize my account and save them, including password, profiles, displays |
| 24 | Analyst | As an Analyst, I want to send request and acquire authorization if needed for any database to reporter |
| 25 | Analyst | As an Analyst, I want to access the data from the reporter |
| 26 | Analyst | As an Analyst, I want toaccess and be able to change the algorithms of the system |
| 27 | Analyst | As an Analyst, I want to access and be able to change the necessary set ups in the system |
| 28 | Analyst | As an Analyst, I want to access the discussion forum and requirement priority list |
| 29 | Analyst | As an Analyst, I want to update the result databases |
| 30 | Analyst | As an Analyst, I want to create basic data dashboards, graphs and visualisations for the executives and supervisors to view so that Manager/Executive can interpret the data easily |

**7. Burndown Chart:**



# 8. Results

Many learning objectives were achieved during this project. The first major objective of applying things learning during the various course to do a complete project was achieved. The project was an industry live project with real company data and requirements from industry. We got to know the real issues faced by companies with regards to data analytics and to come up with problem statement which would be addressed by the project was a huge take away. The opportunity to speak with industry experts and gain knowledge in the field and have a customer relationship was an incredible opportunity. The project could be adopted by the industry with the focus being able to fulfill customer requirements with a flexible and easy to use dashboard was a driving force in working towards a good product.

The software system design and implementation course provided me the opportunity to employ the skills acquired through the various course work I did in MSIT program. My responsibilities included project management, scope refinement, time management, assigning tasks. I took up the challenge of data cleaning and analysis of attrition data. It was fulfilling when I was able to provide actionable inputs from the data. The attrition data based on gender, age group, job type and locations would help the company understand areas of concern and issues. Which could be used for corrective actions in specific fields.

We were able to find that the attrition was maximum with employees who were in the company for less than 1 year, the forecast of female employees leaving the organization is in a downward trend. the forecast of male employee leaving is in a slight upward trend. the overall trend is of increasing number of employees leaving the organization. up to the age of 35 there is an estimated reduction in the employee attrition rate. from 36 up to 55 years there is a predicted increase in the employee attrition rate. Following are some of the screenshots of data displayed on the dashboard.

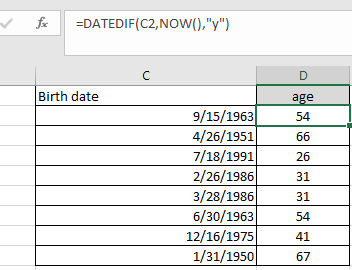
Data manipulation operations on the Excel.

The following were the data manipulation operations done on the excel.

1) Calculating Age from Date of birth.

**=DATEDIF(C2,NOW(),"y")** : in place of C2 provide the column with the Birth date.

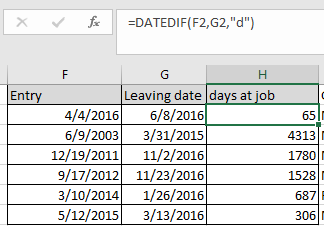
Figure 4.1 Age Calculation Formula



Calculate number of days in the company.

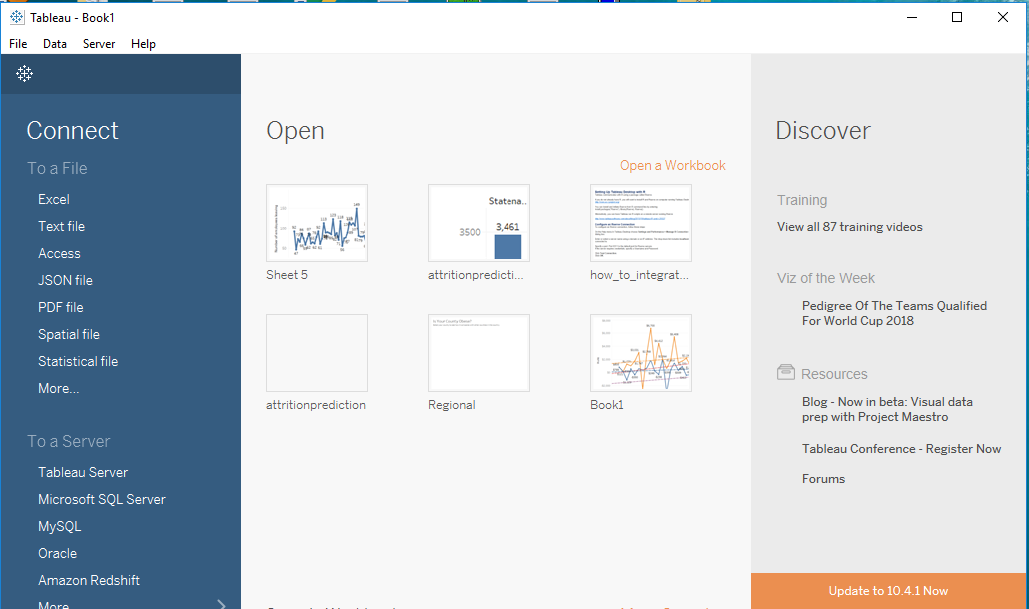
**=DATEDIF(F2,G2,"d") :** in place of F2 and G2 provide start and end date columns.

Figure 4.2 Age Calculation Results



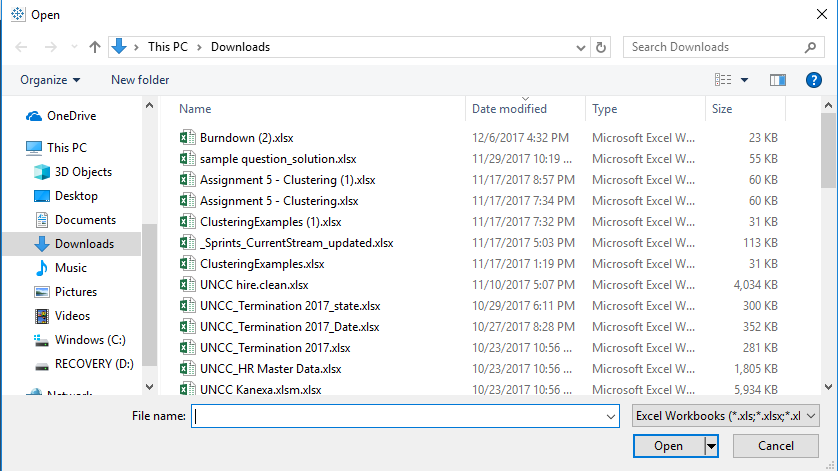
Operations using Tableau.

Figure 4.3 Excel Data Import in Tableau



Select Excel

Figure 4.4 Excel Data Selection in Tableau



Select appropriate file to open.

Figure 4.5 Excel File Selection

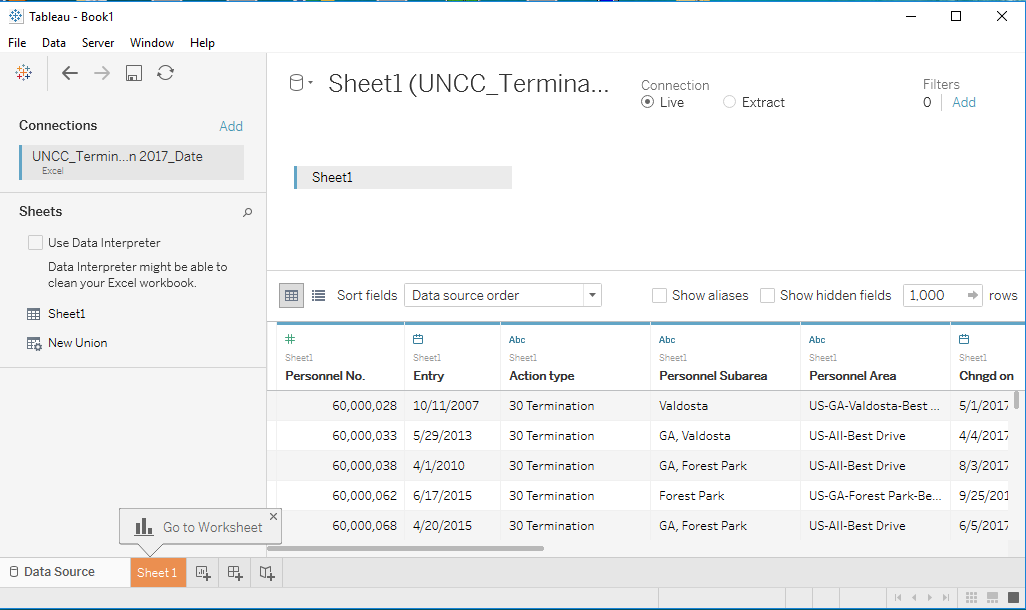
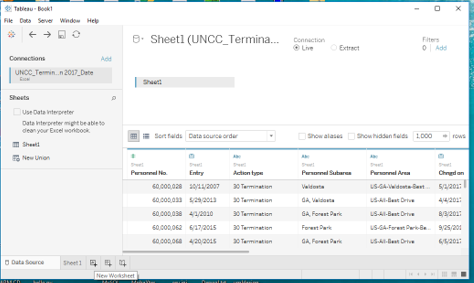
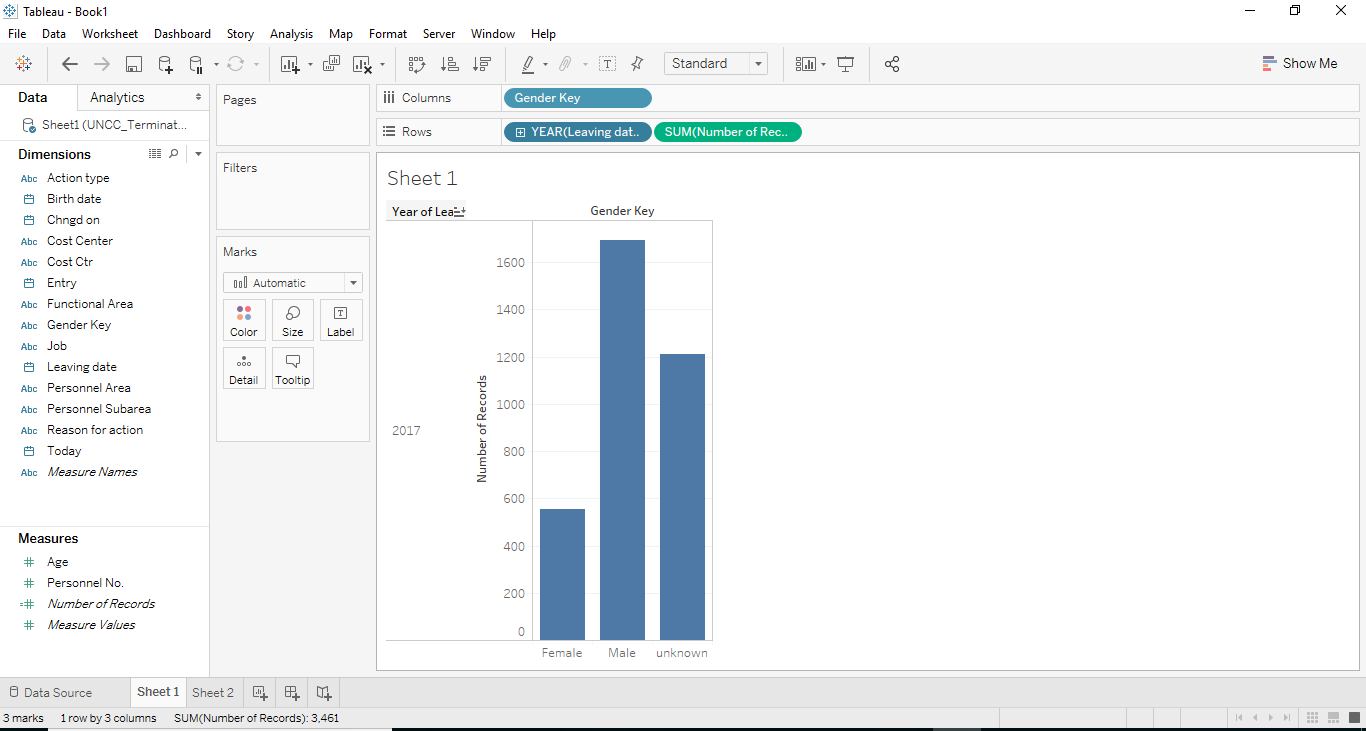


Figure 4.6 New Worksheet Selection in Tableau

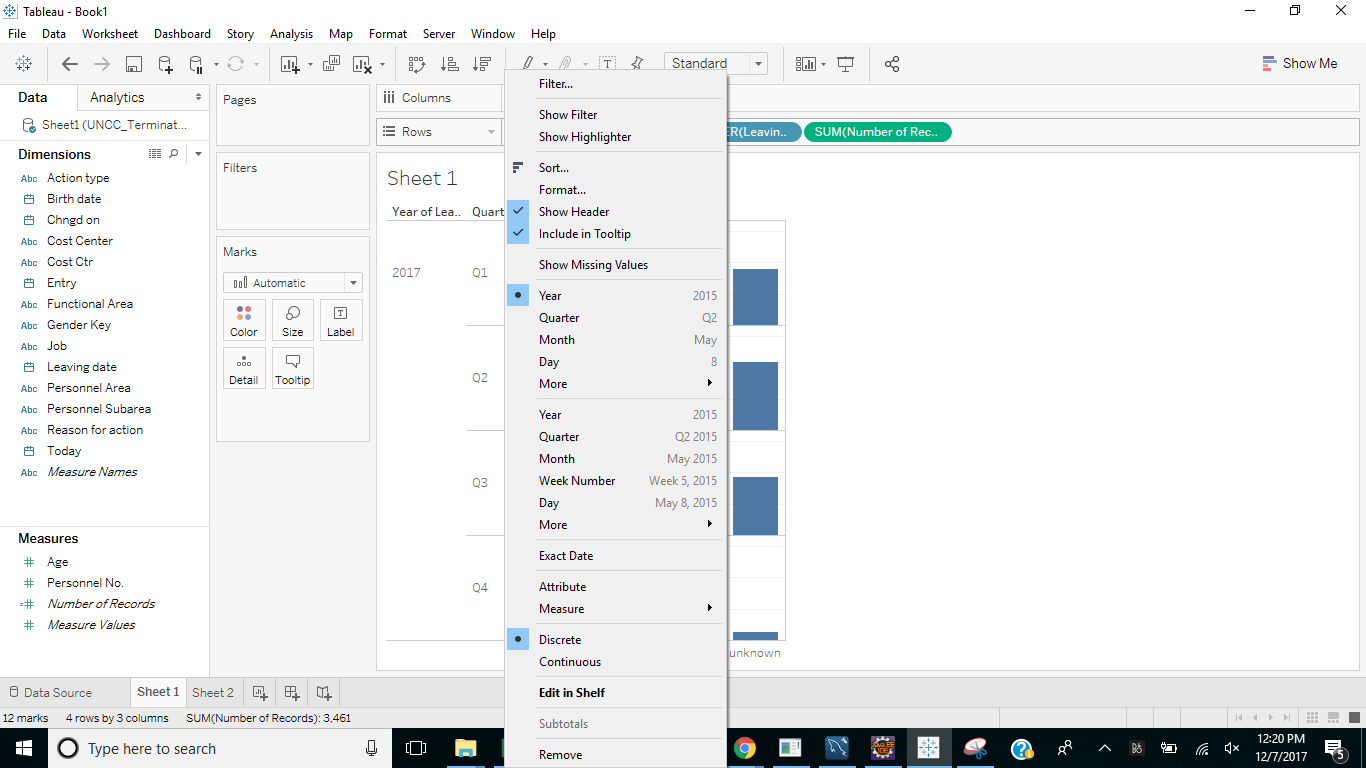


Select new worksheet to work on the visualization.

Figure 4.7 Tableau Bar Chart****

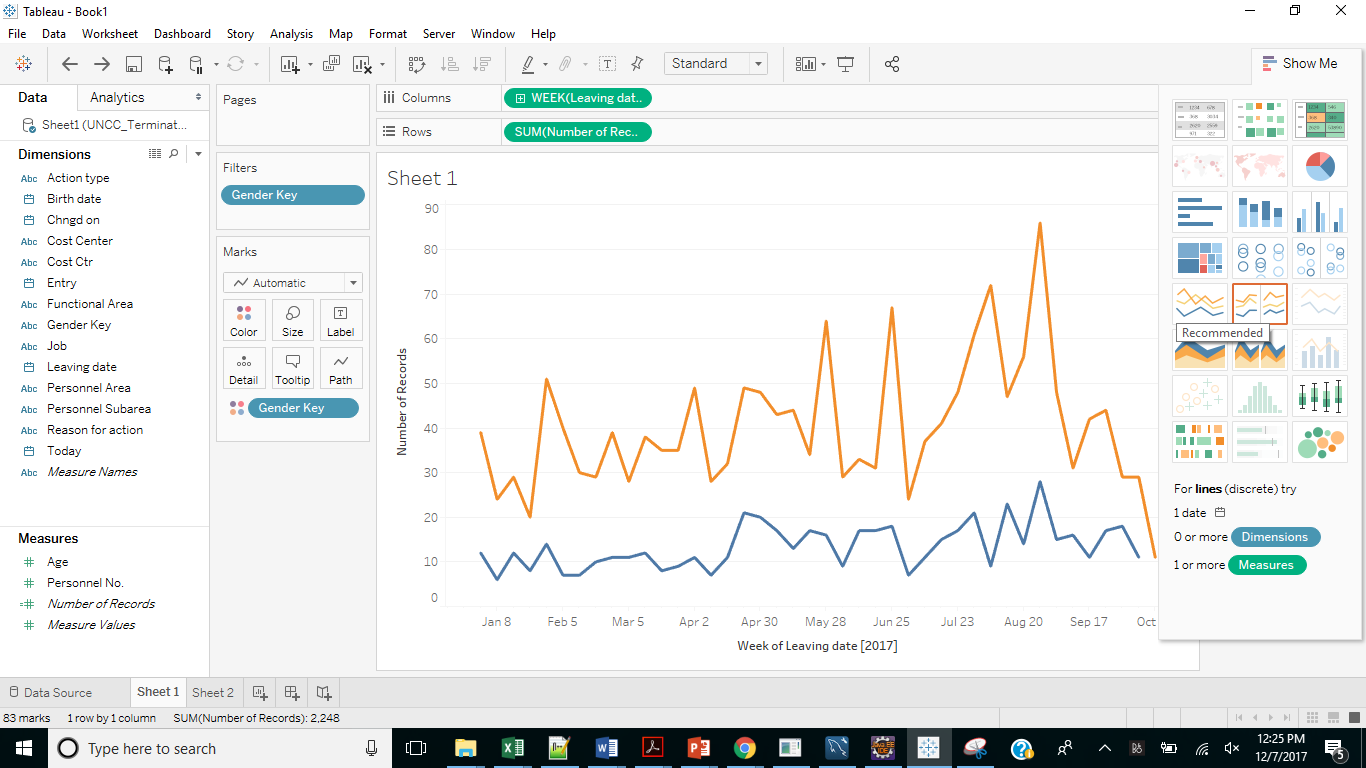
Add data from dimensions and measures. For the forcast of people attrition. Add Gender key to columns, leaving date to rows and select number of records from measures.

Figure 4.8 Frequency Selection in Tableau

****

Click on Years and from the drop down select the frequency for which you need the output. In our case we select weeks. Add gender to filter to remove the unknown members from the gender column.

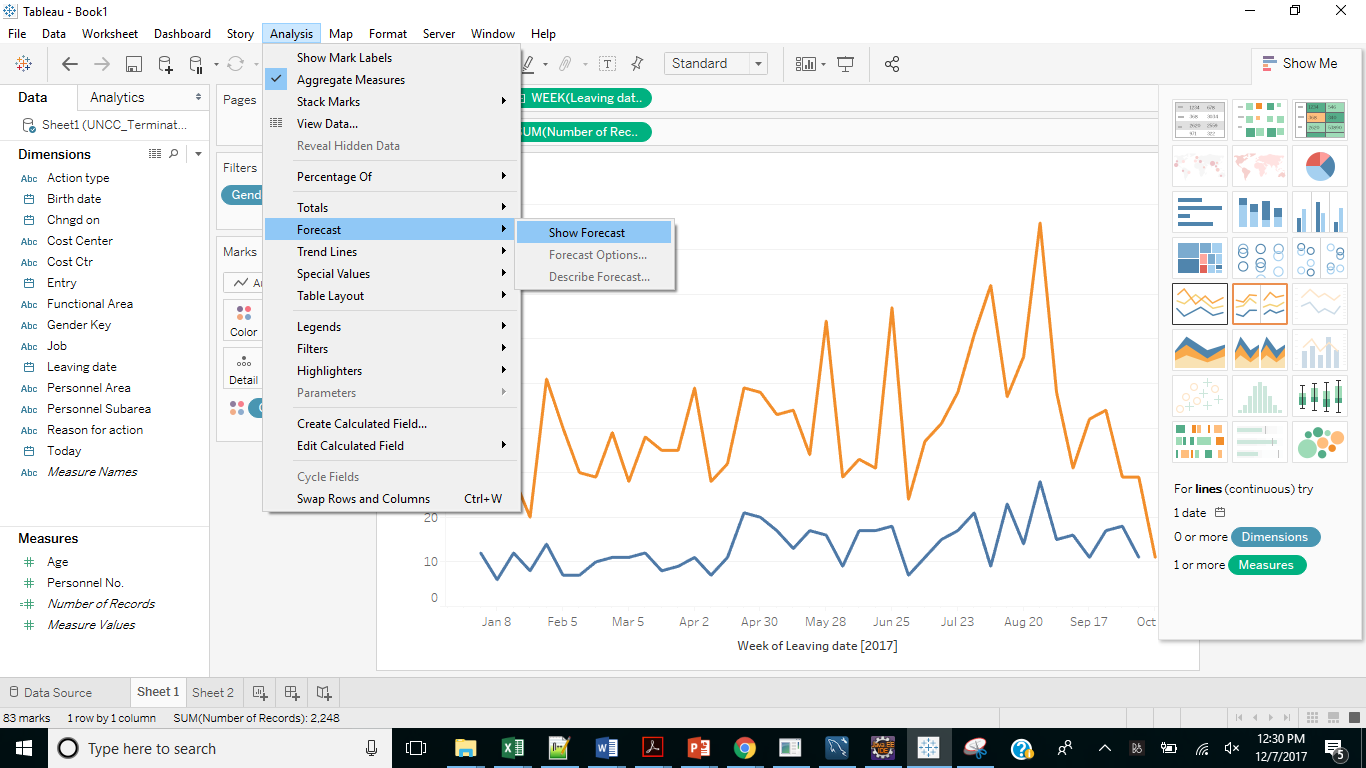
Figure 4.9 Tableau Line Graph



Select appropriate graph for showcasing the data. We use line graph as its best for forcasting.

Select Analysis -> Forcast -> show forcast

Figure 4.10 Tableau Forecast



Select Analysis -> Forcast -> Forcast options

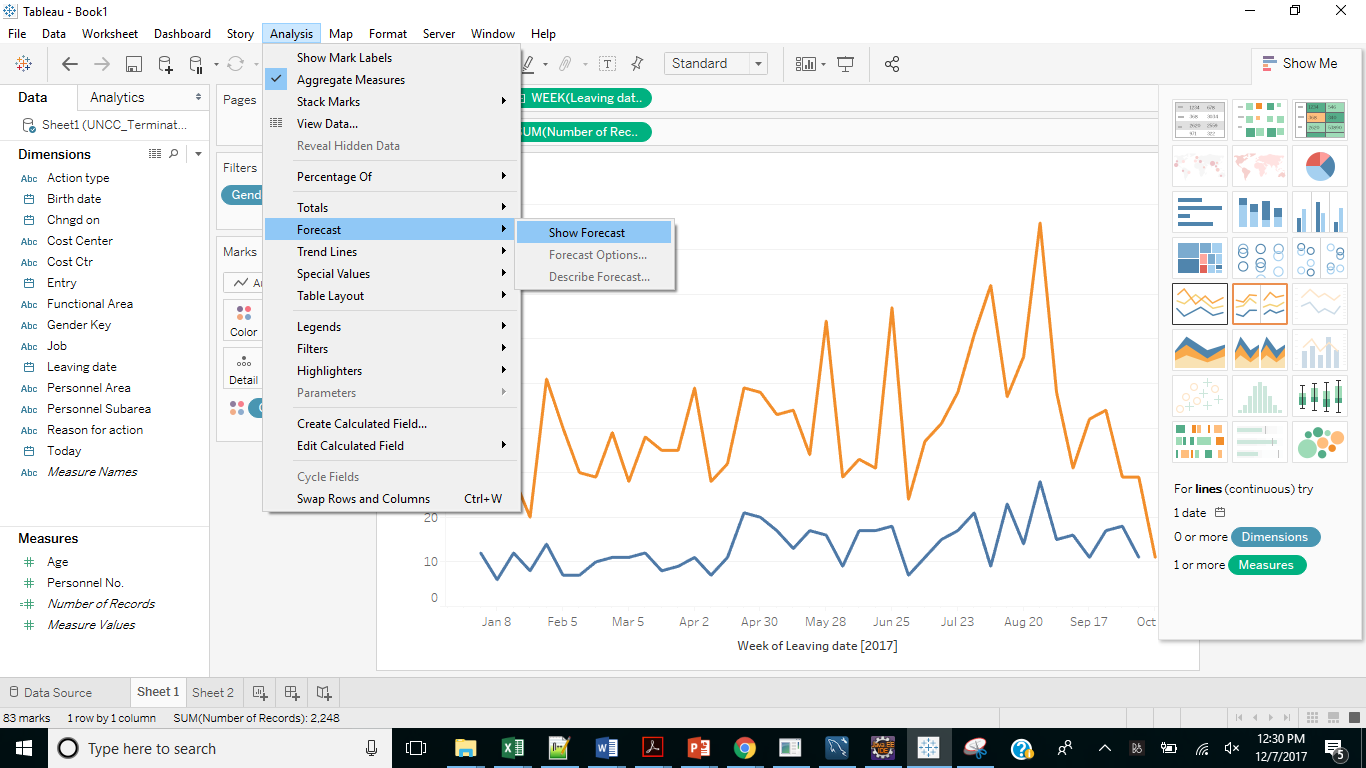
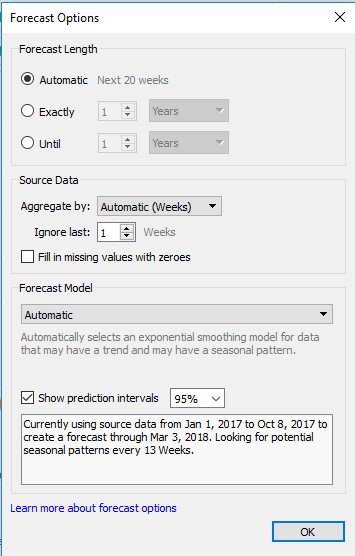
Figure 4.11 Show Forecast Selection****

Figure 4.12 Forecast Options



Under forecast options select the appropriate model and period.

Figure 4.13 Model and Time Selection

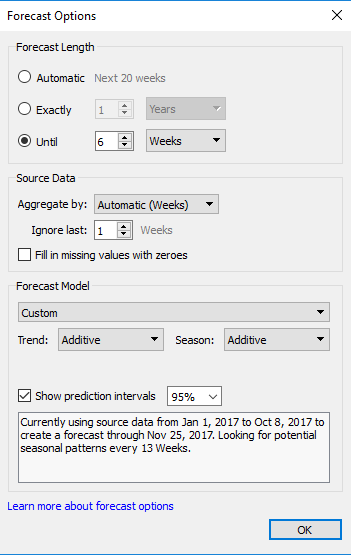
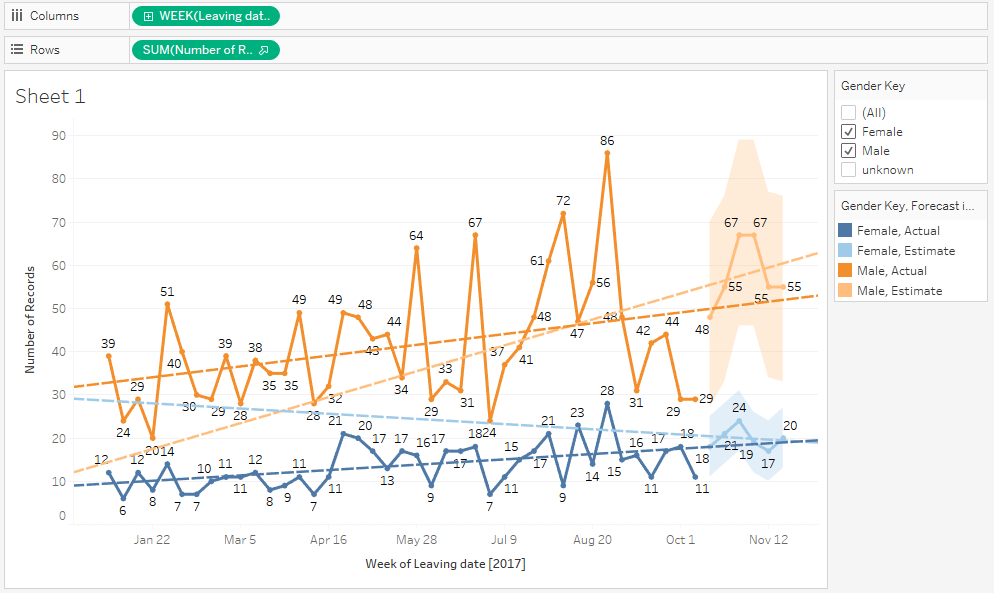


Figure 4.14 Trend Line Selection in Tableau



Right click on the data to show trend lines.

This data can now be exported to excel using the following procedure.

Select worksheet -> export -> crosstab to excel

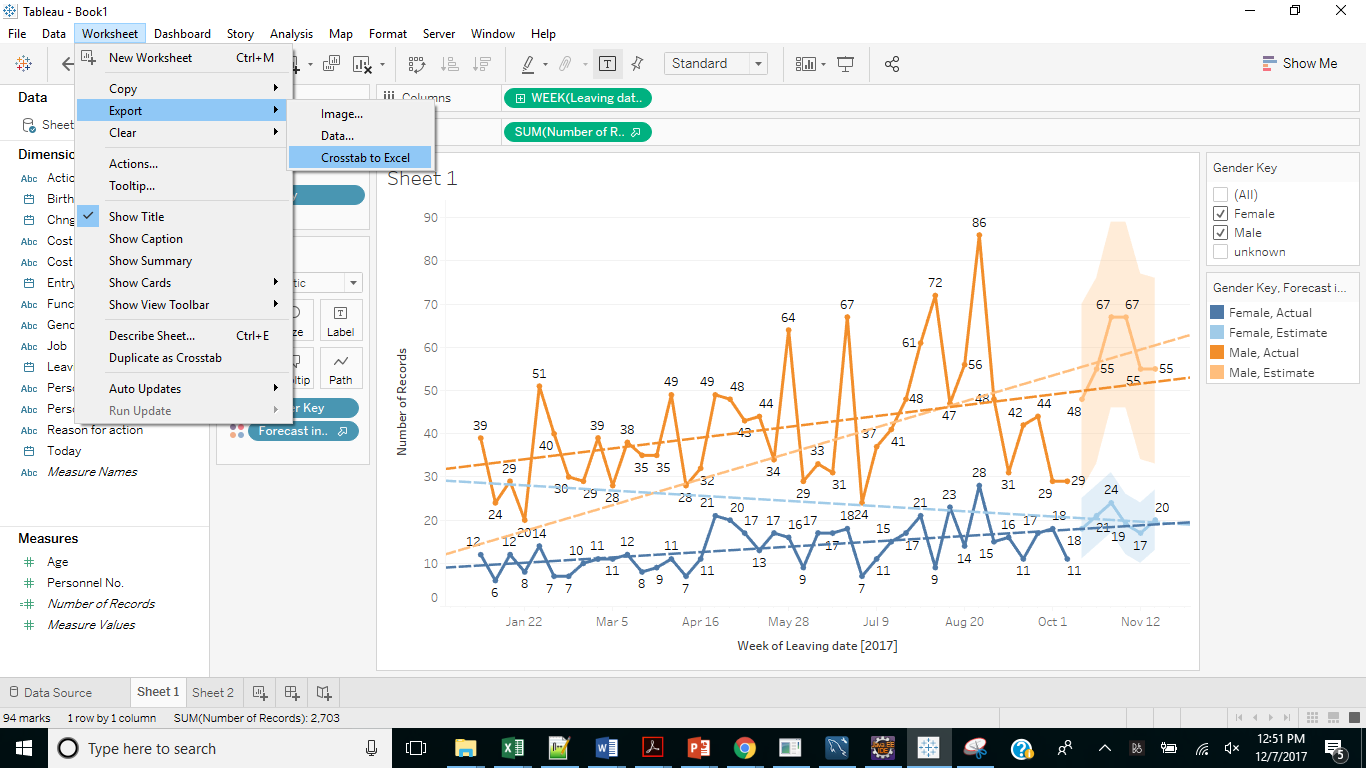
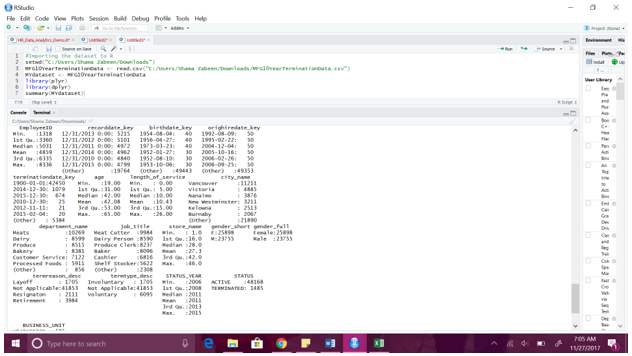
Figure 4.15 Cross to Excel in Tableau****

Figure 4.16 Data Exported into Excel with Labels



The data is exported into excel with the following labels.

The data set being used here is an employee data set containing data for reasons of leaving and job title, location, age, gender, position, date of leaving, tenure of work with the company and many more for a period of 2006-2015. Making use of the R script, we have trained the data set and tested the trained features to predict the reasons and factors affecting the attrition rate. A detailed note on the project execution is given below with step wise description of the pseudo code and the output screenshots of the visualizations made.

Figure 4.17 Data Loading in R****

It is always useful to have an idea on the percent/proportion of the terminations of the entire employee population.

**Pseudo code:**

**StatusCount<- as.data.frame.matrix(MYdataset %>%**

**group\_by(STATUS\_YEAR) %>%**

**select(STATUS) %>%**

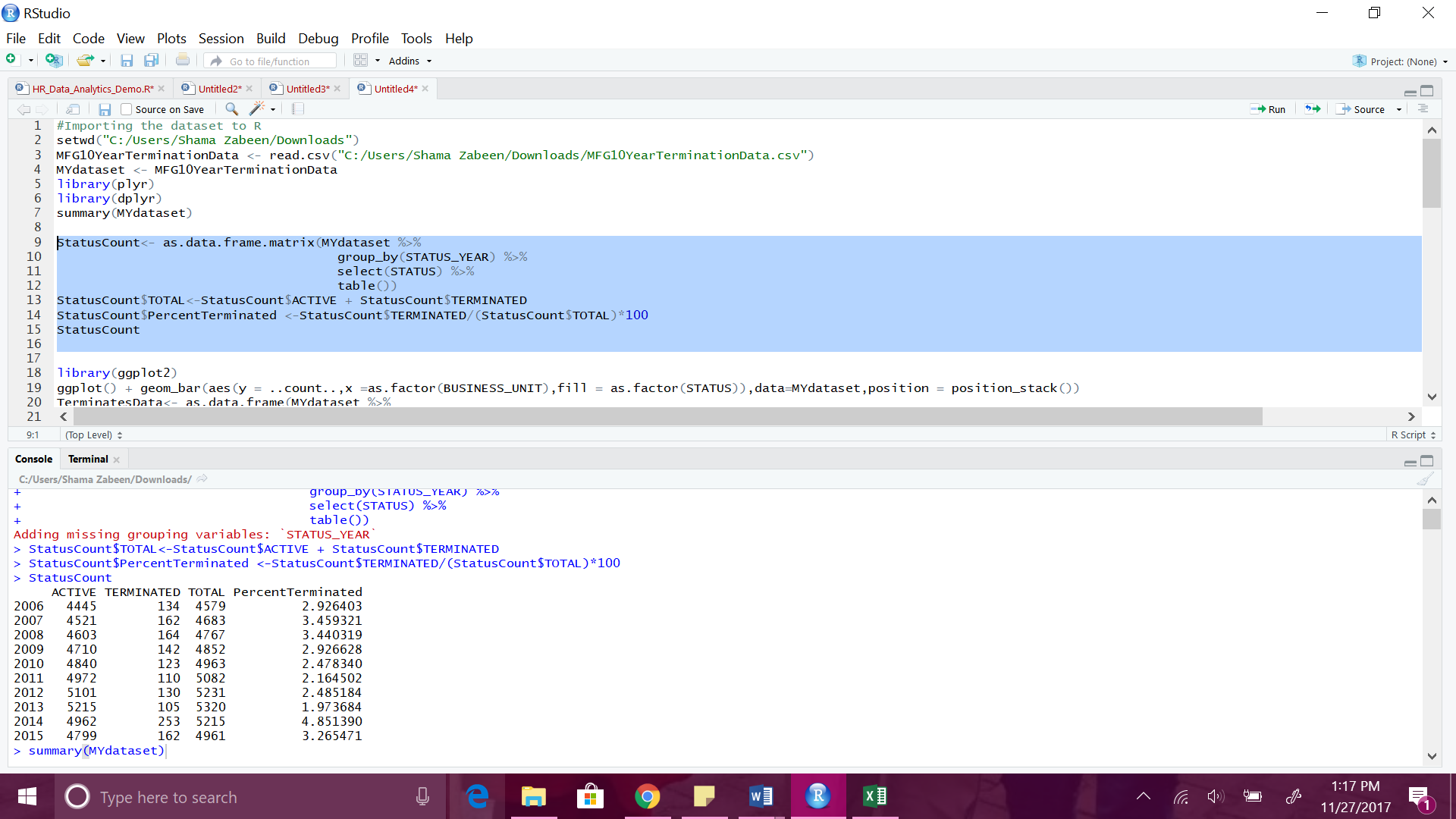
**table())**

**StatusCount$TOTAL<-StatusCount$ACTIVE + StatusCount$TERMINATED**

**StatusCount$PercentTerminated <-StatusCount$TERMINATED/(StatusCount$TOTAL)\*100**

**StatusCount**

Figure 4.18 Percentage of Employees Leaving the Company



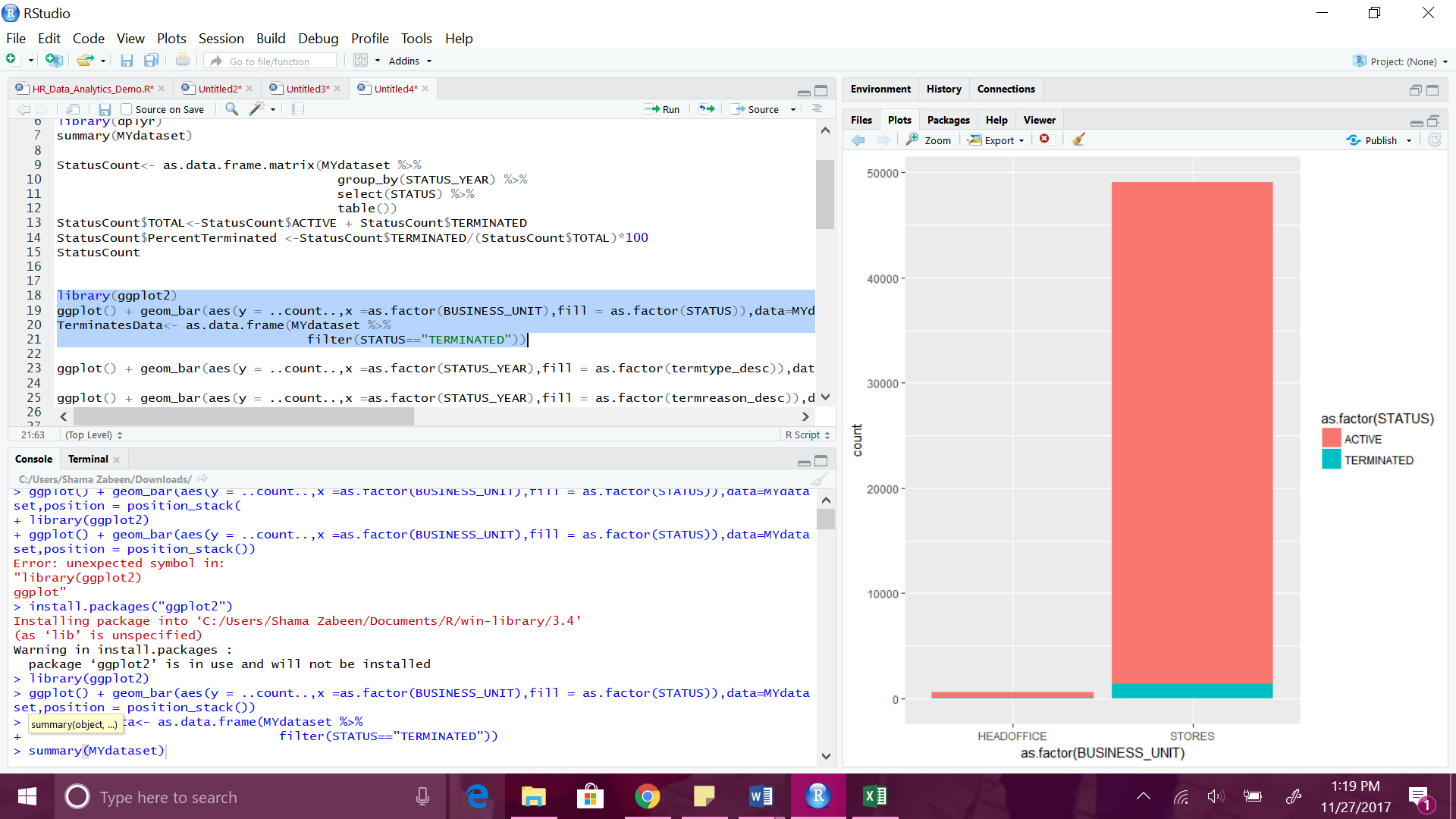
To find out where the terminations are occurring the most, lets have a look at few graphs:

**BY BUSINESS UNIT:**

**library(ggplot2)**

**ggplot() + geom\_bar(aes(y = ..count..,x =as.factor(BUSINESS\_UNIT),fill = as.factor(STATUS)),data=MYdataset,position = position\_stack())**

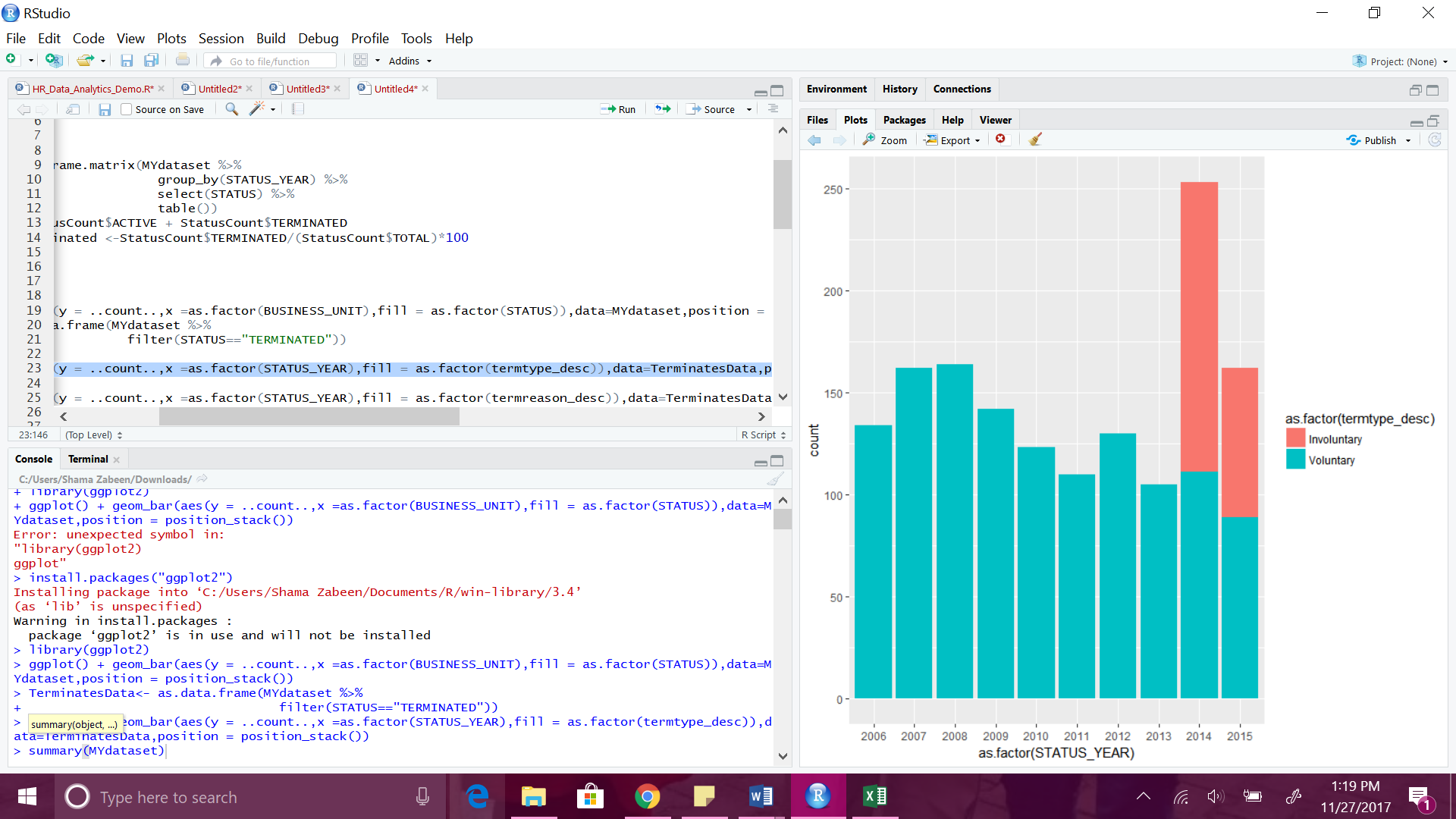
Figure 4.19 Business Unit Based Attrition

****

The terminations in the last 10 years have predominantly occurred in the stores business units, and only one termination was seen in the HR Technology at the head office.

**BY TERMINATION TYPE AND STATUS YEAR:TerminatesData<- as.data.frame(MYdataset %>% filter(STATUS=="TERMINATED"))ggplot() + geom\_bar(aes(y = ..count..,x =as.factor(STATUS\_YEAR),fill = as.factor(termtype\_desc)),data=TerminatesData,position = position\_stack())**

Figure 4.20 Termination Type and Status Year

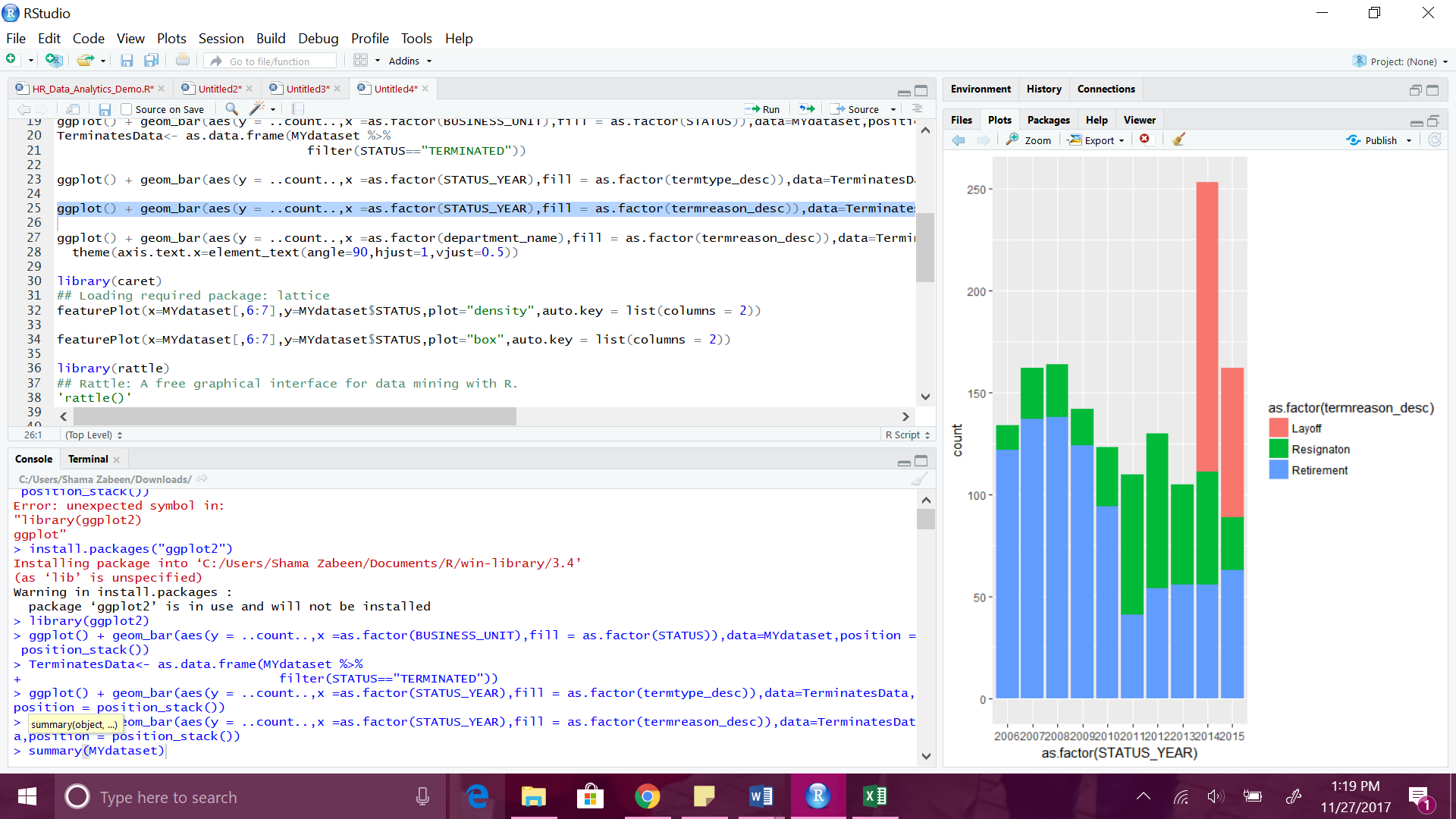
****

Most of the terminations so far seem to be voluntary year by year, except for in the most recent years where some involuntary terminations were noticed.

**TERMINATION BY STATUS YEAR AND THE REASON:**

**ggplot() + geom\_bar(aes(y = ..count..,x =as.factor(STATUS\_YEAR),fill = as.factor(termreason\_desc)),data=TerminatesData,position = position\_stack())**

Figure 4.21 Termination by Status Year and Reason

****

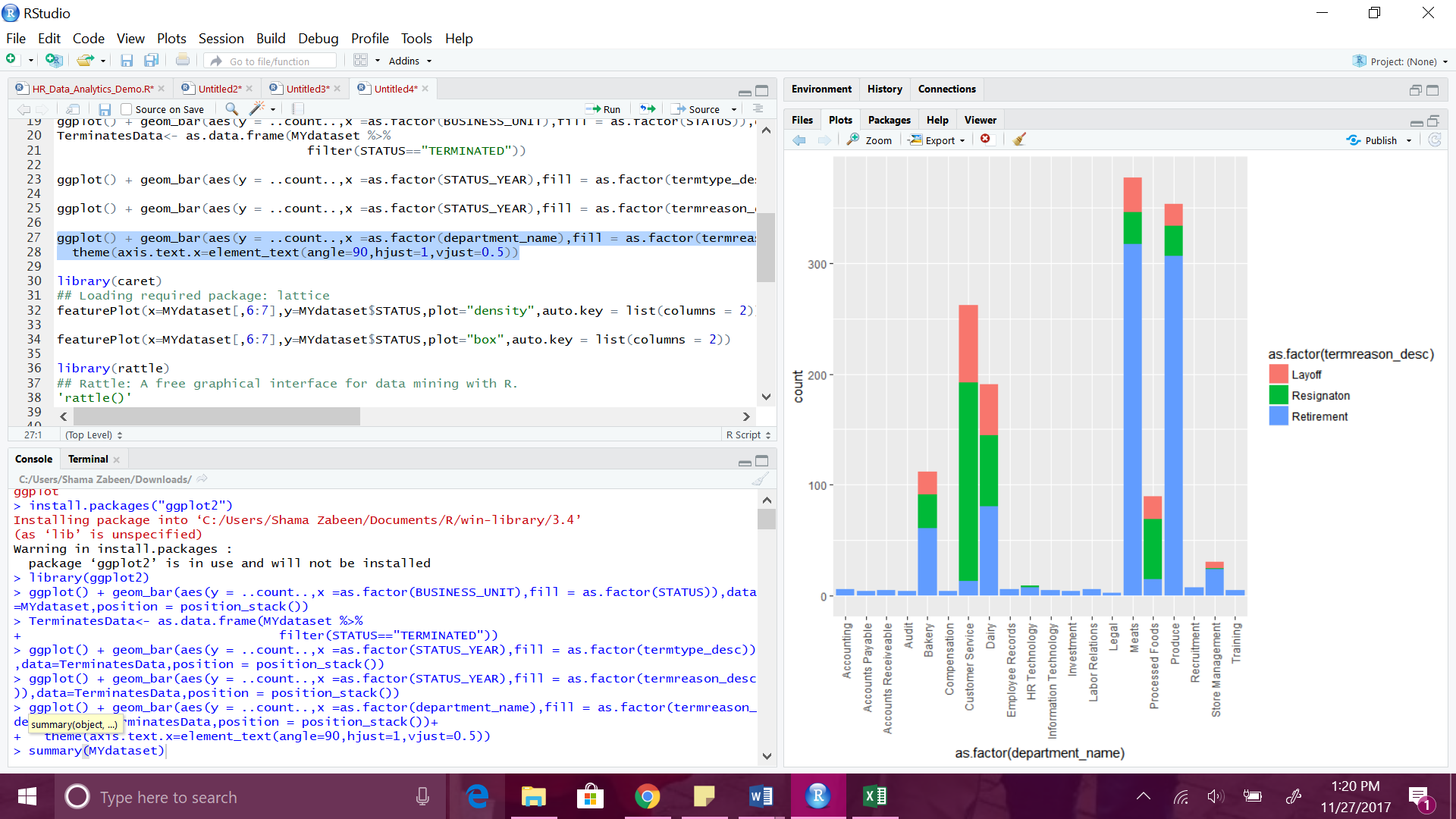
Layoffs were noticed in 2014 and 2015 accounted for the involuntary terminations.

**TERMINATION BY REASON AND DEPARTMENT:**

**ggplot() + geom\_bar(aes(y = ..count..,x =as.factor(department\_name),fill = as.factor(termreason\_desc)),data=TerminatesData,position = position\_stack())+**

**theme(axis.text.x=element\_text(angle=90,hjust=1,vjust=0.5))**

Figure 4.22 Termination by Reason and Department

****

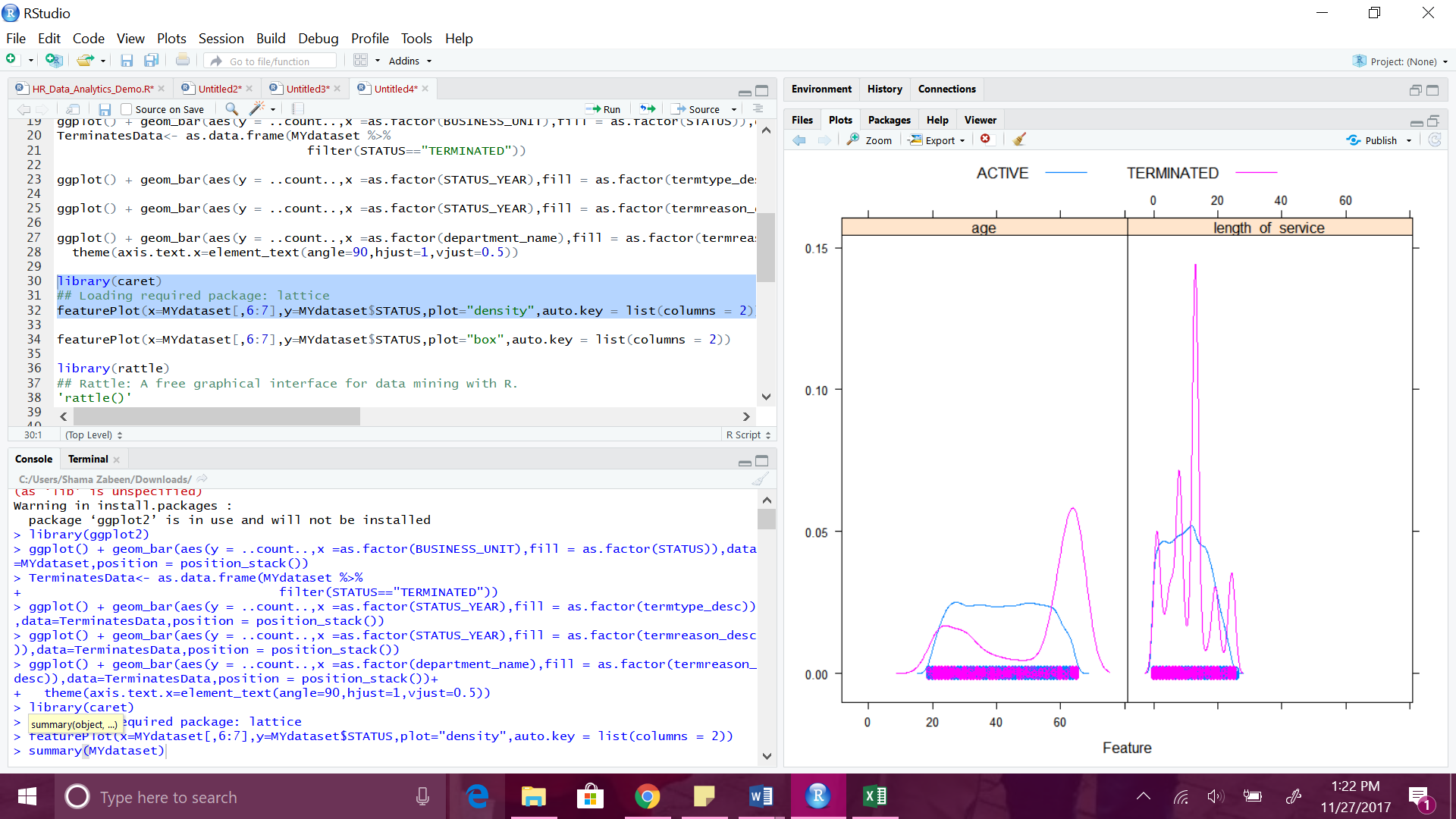
Customer Service has a much larger proportion of the resignation compared to other departments. Also, retirement in general is high in all the departments.

**TERMINATION BY AGE AND DEPARTMENT:**

**library(caret)**

**featurePlot(x=MYdataset[,6:7],y=MYdataset$STATUS,plot="density",auto.key = list(columns = 2))**

Figure 4.23 Termination by Age and Department

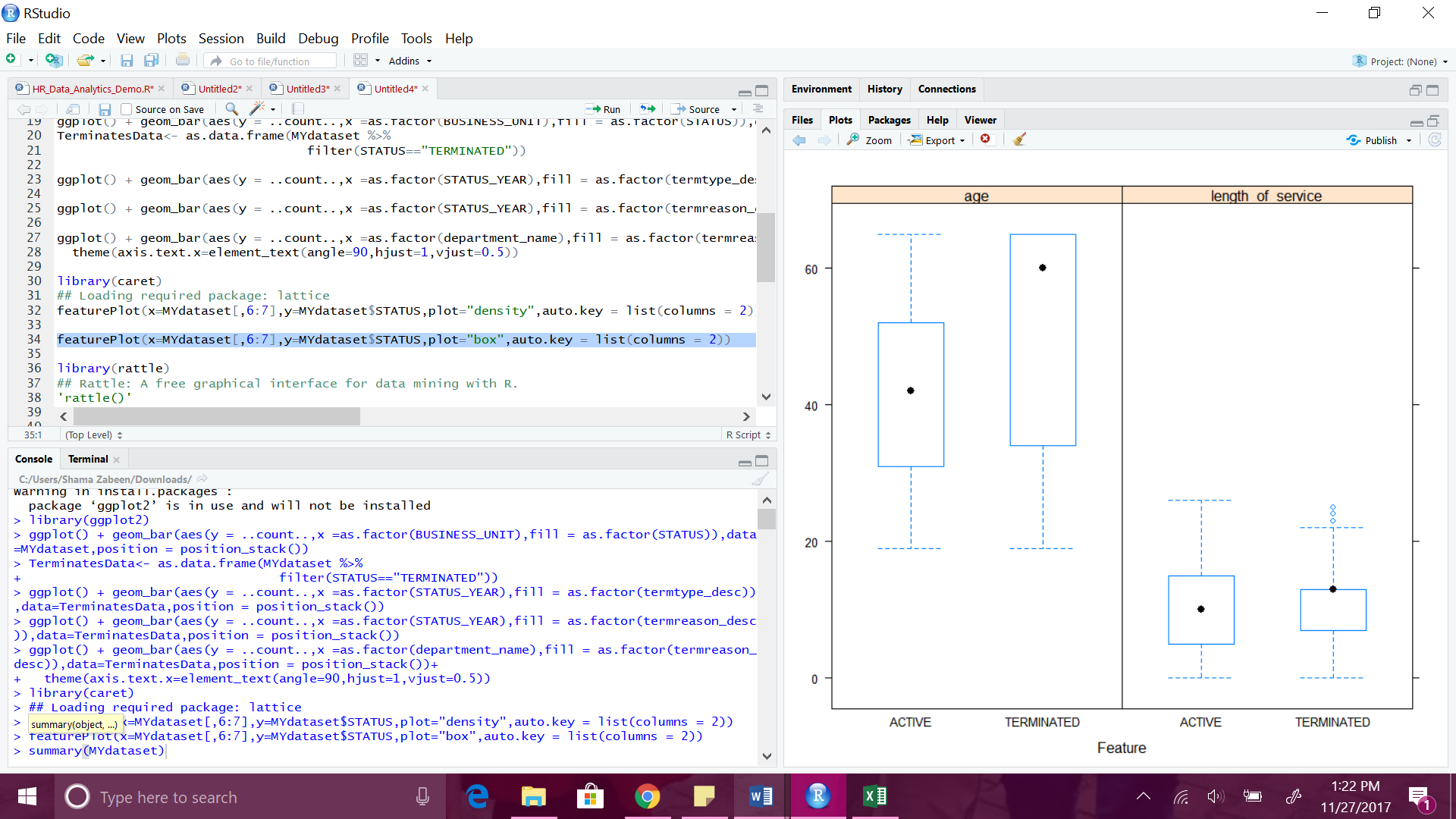
****

For termination there is some elevation from 20 to 30 and a spike at 60. For the length of service there are 5 spikes, one around 1 year and the other around 5 years, and a big one around 15 years, and a couple at 20 and 25 years.

**AGE AND LENGTH OF SERVICE DISTRIBUTIONS BY STATUS:**

**featurePlot(x=MYdataset[,6:7],y=MYdataset$STATUS,plot="box",auto.key = list(columns = 2))**

Figure 4.24 Age and Length of Service Distribution by Status

****

Box plots show high average age for terminations as compared to active employee. Length of the service shows not much difference between the active and the terminated.

**TRAINING AND TESTING OF THE DATA:**

We have 10 years of historical data, of which we are making use of 9 years of data for training the model, and the 10th year data to test the trained model. Moreover, we will use 10 fold cross validation on the training data.

**Partitioning the data sets:**

**library(rattle)**

**'rattle()'**

**library(magrittr) # For the %>% and %<>% operators.**

**building TRUE**

**scoring # A pre-defined value is used to reset the random seed so that results are repeatable.**

**crv$seed 42**

**MFG10YearTerminationData read.csv("C:/Users/Shama Zabeen/Downloads/MFG10YearTerminationData.csv")**

**MYdataset #Create training and testing datasets**

**set.seed(crv$seed)**

**MYnobs nrow(MYdataset) # 52692 observations**

**MYsample subset(MYdataset,STATUS\_YEAR<=2014)**

**MYvalidate NULL**

**MYtest subset(MYdataset,STATUS\_YEAR== 2015)**

**MYinput c("age", "length\_of\_service", "gender\_full", "STATUS\_YEAR", "BUSINESS\_UNIT")**

**MYnumeric c("age", "length\_of\_service", "STATUS\_YEAR")**

**MYcategoric c("gender\_full", "BUSINESS\_UNIT")**

**MYtarget "STATUS"**

**MYrisk NULL**

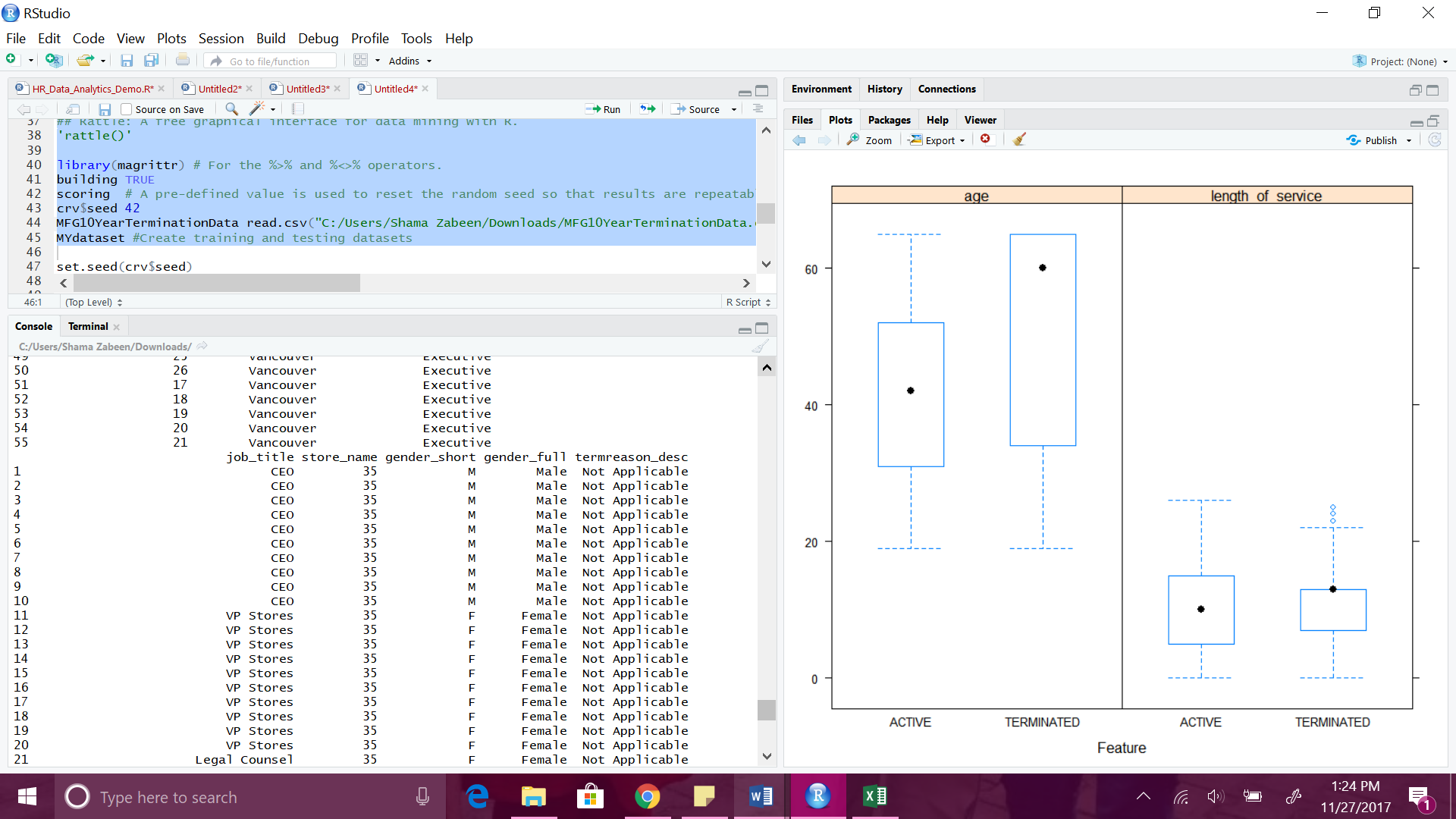
**MYident "EmployeeID"**

**MYignore c("recorddate\_key", "birthdate\_key", "orighiredate\_key", "terminationdate\_key", "city\_name", "gender\_short", "termreason\_desc", "termtype\_desc","department\_name","job\_title", "store\_name")**

**MYweights NULL  
MYTrainingData<-MYtrain[c(MYinput, MYtarget)]**

**MYTestingData<-MYtest[c(MYinput, MYtarget)]**

Figure 4.25 Partitioning the Dataset

****

**Decision Tree:**

**library(rattle)**

**library(rpart, quietly=TRUE)**

**set.seed(crv$seed)**

**MYrpart = rpart(STATUS ~ .,**

**data=MYtrain[, c(MYinput, MYtarget)],**

**method="class",**

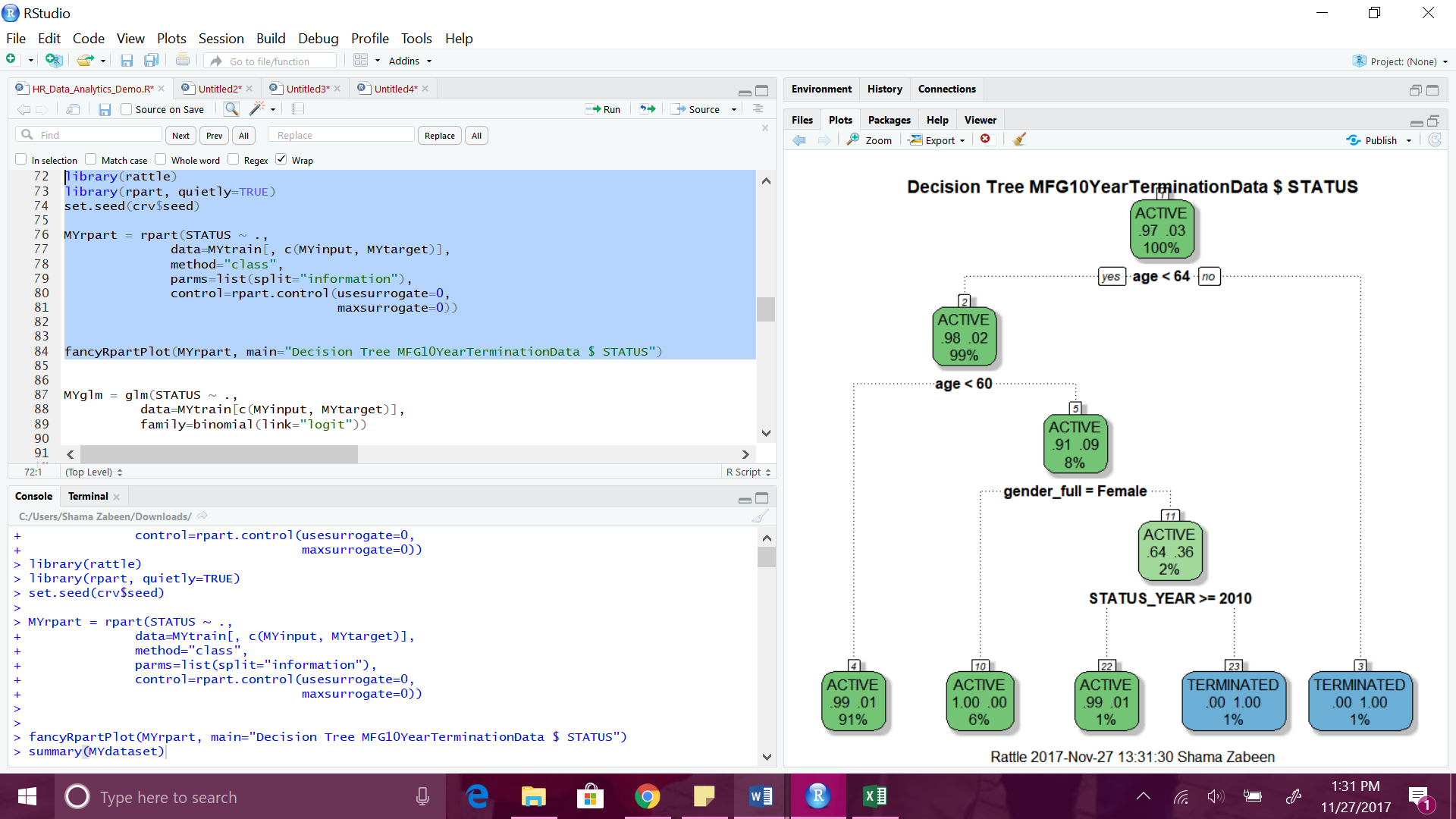
**parms=list(split="information"),**

**control=rpart.control(usesurrogate=0,**

**maxsurrogate=0))**

**fancyRpartPlot(MYrpart, main="Decision Tree MFG10YearTerminationData $ STATUS")**

Figure 4.26 Decision Tree

****

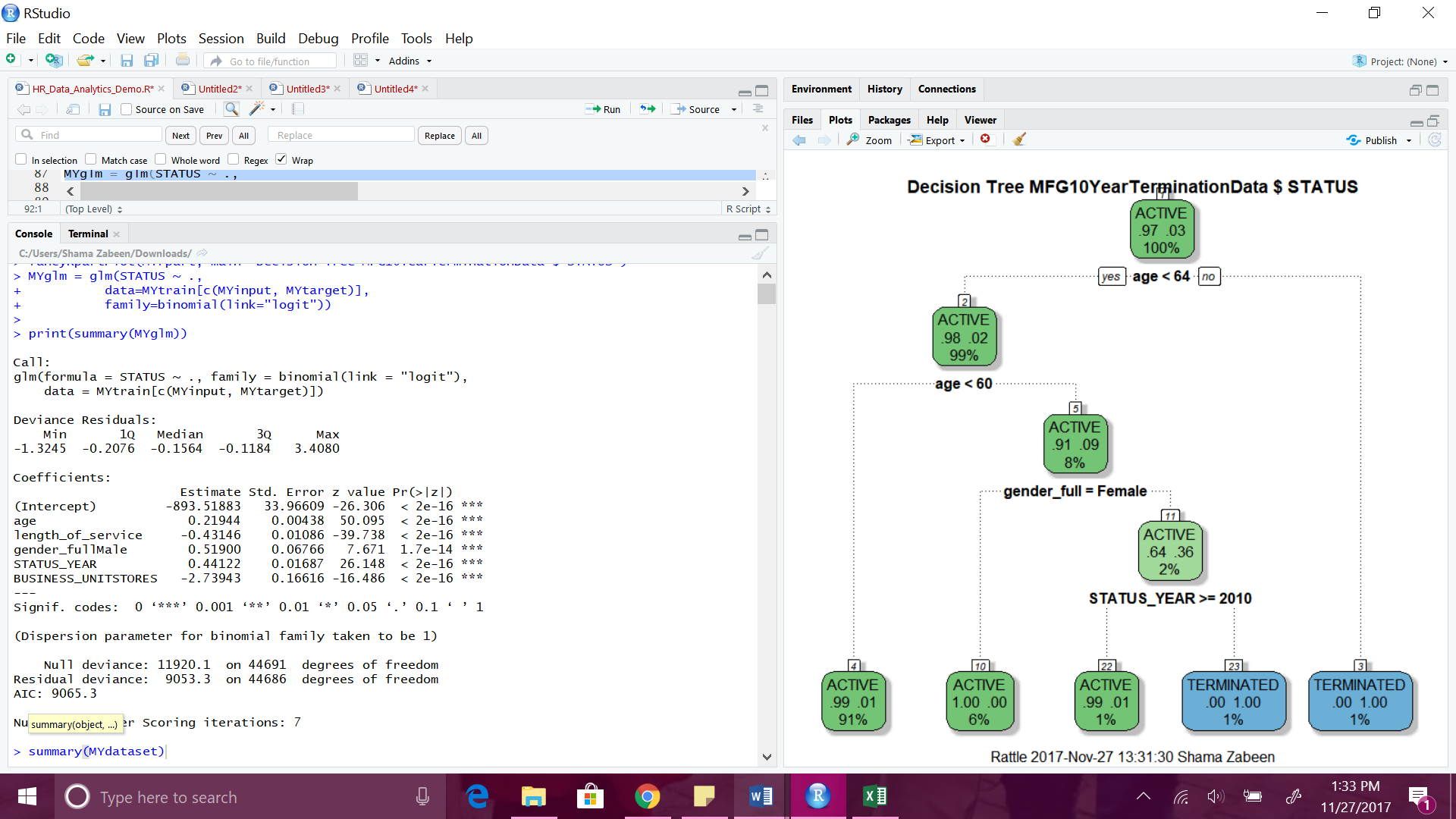
**MYglm = glm(STATUS ~ .,**

**data=MYtrain[c(MYinput, MYtarget)],**

**family=binomial(link="logit"))**

**print(summary(MYglm))**

Figure 4.27 Data Training

****

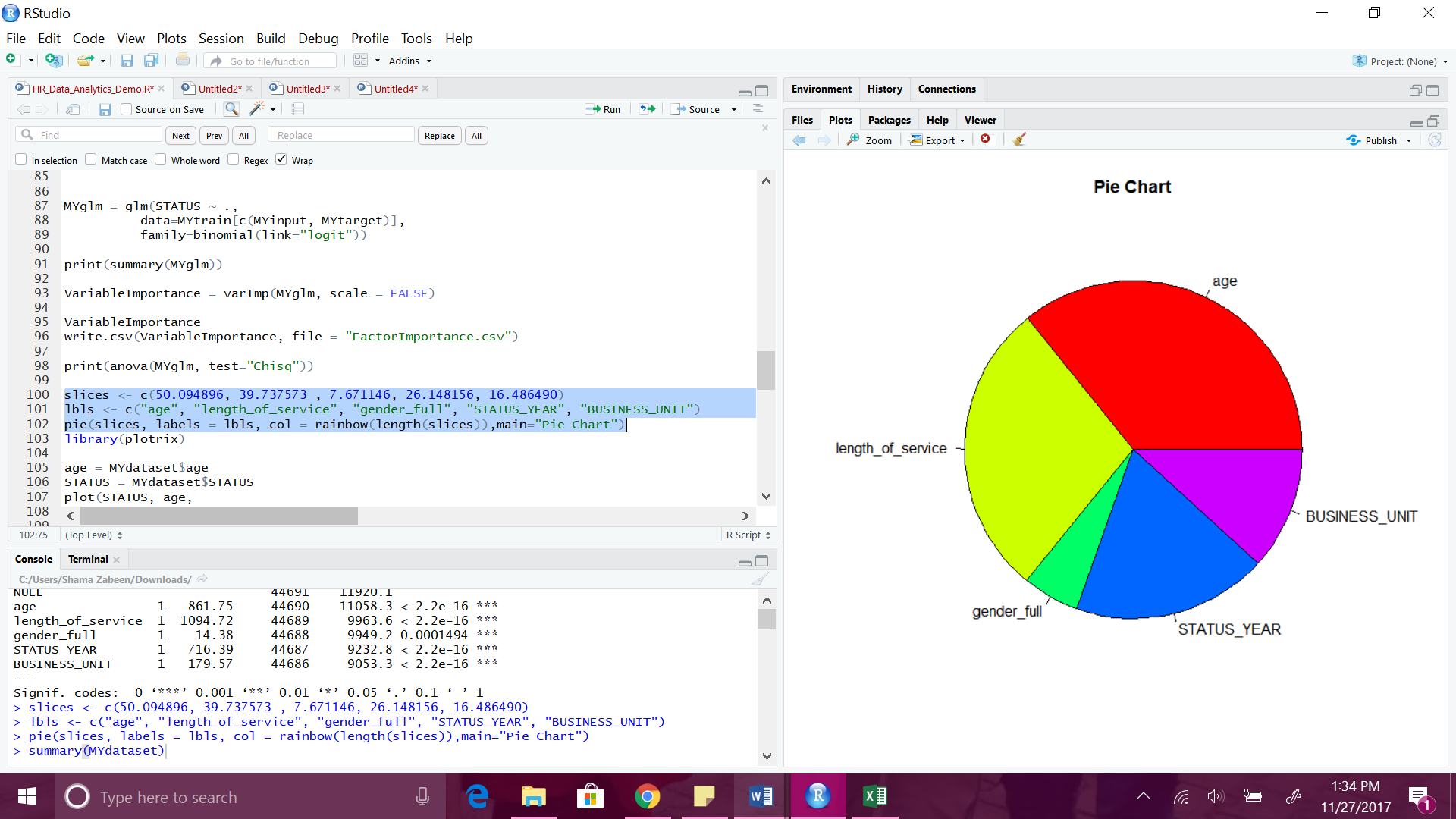
**PIE CHART FOR THE FACTORS:**

**slices <- c(50.094896, 39.737573 , 7.671146, 26.148156, 16.486490)**

**lbls <- c("age", "length\_of\_service", "gender\_full", "STATUS\_YEAR", "BUSINESS\_UNIT")**

**pie(slices, labels = lbls, col = rainbow(length(slices)),main="Pie Chart")**

Figure 4.28 Pie Chart for the Factors



**TESTING THE TRAINED DATA BASED ON STATUS AND AGE:**

**library(plotrix)**

**age = MYdataset$age**

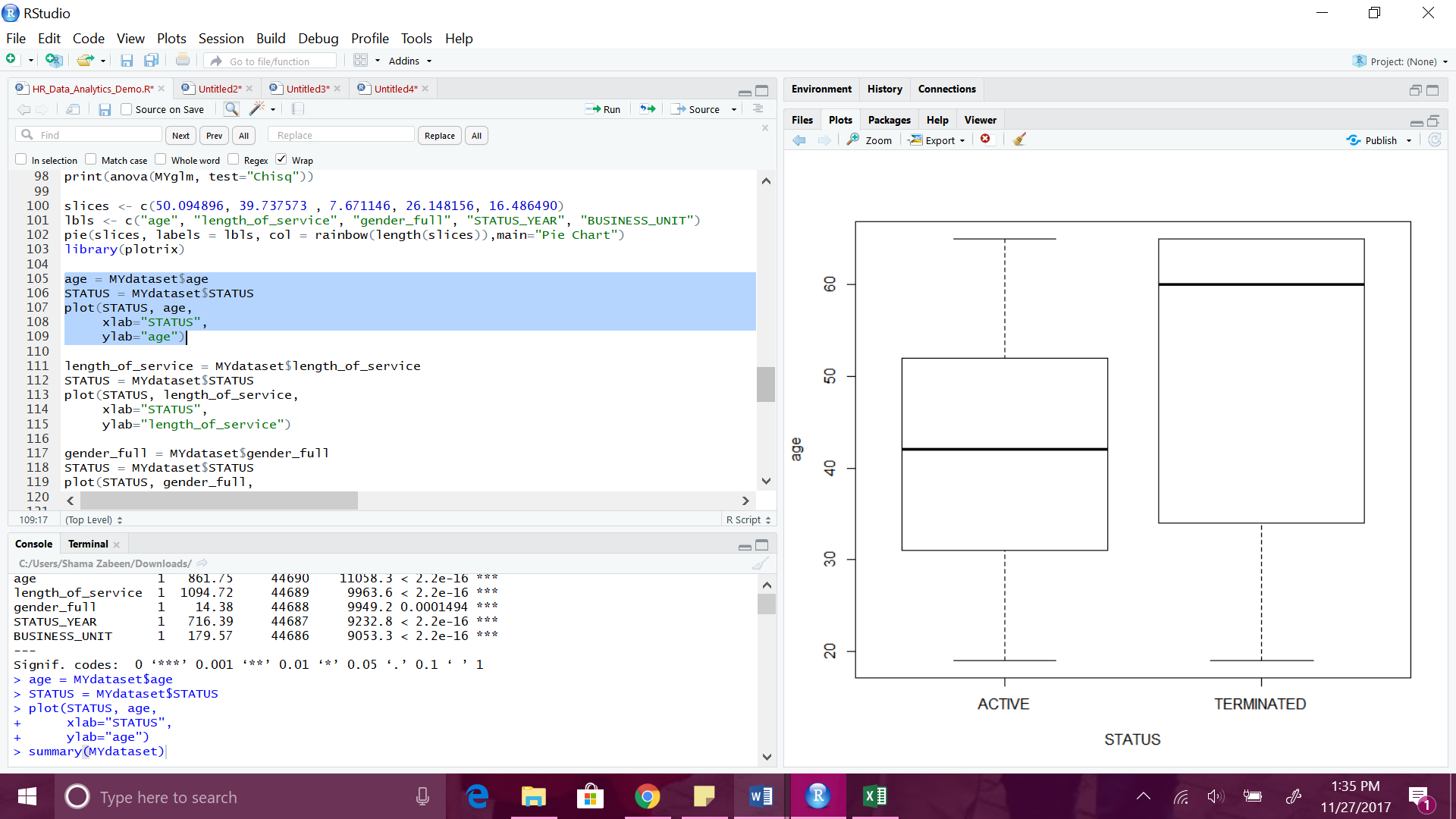
**STATUS = MYdataset$STATUS**

**plot(STATUS, age,**

**xlab="STATUS",**

**ylab="age")**

Figure 4.29 Testing the Trained Data Based on Status and Age

****

**TESTING THE TRAINED DATA BASED ON STATUS AND LENGTH OF SERVICE:**

**length\_of\_service = MYdataset$length\_of\_service**

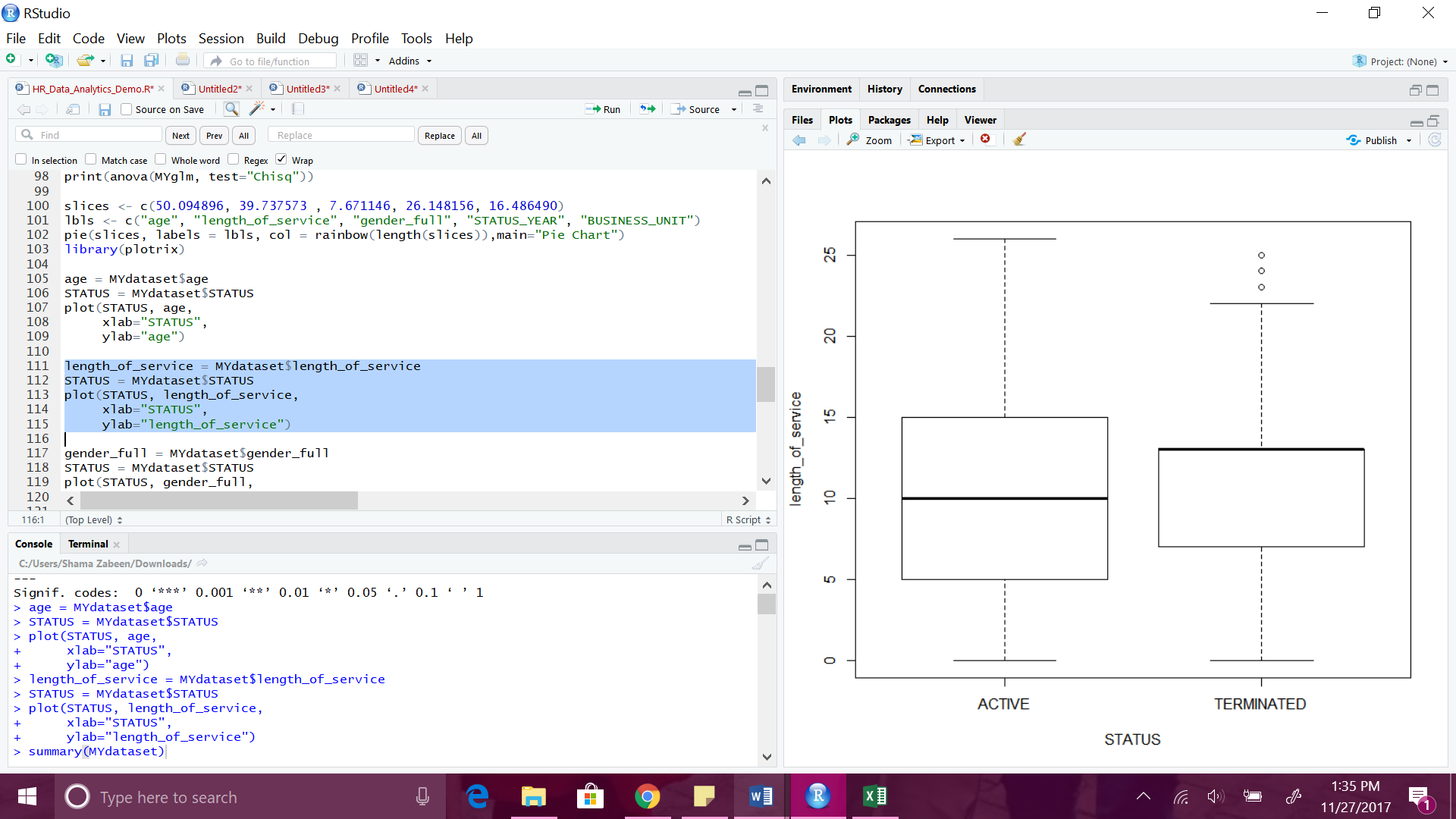
**STATUS = MYdataset$STATUS**

**plot(STATUS, length\_of\_service,**

**xlab="STATUS",**

**ylab="length\_of\_service")**

Figure 4.30 Based on Length of Service

****

**TESTING THE TRAINED DATA BASED ON STATUS AND GENDER:**

**gender\_full = MYdataset$gender\_full**

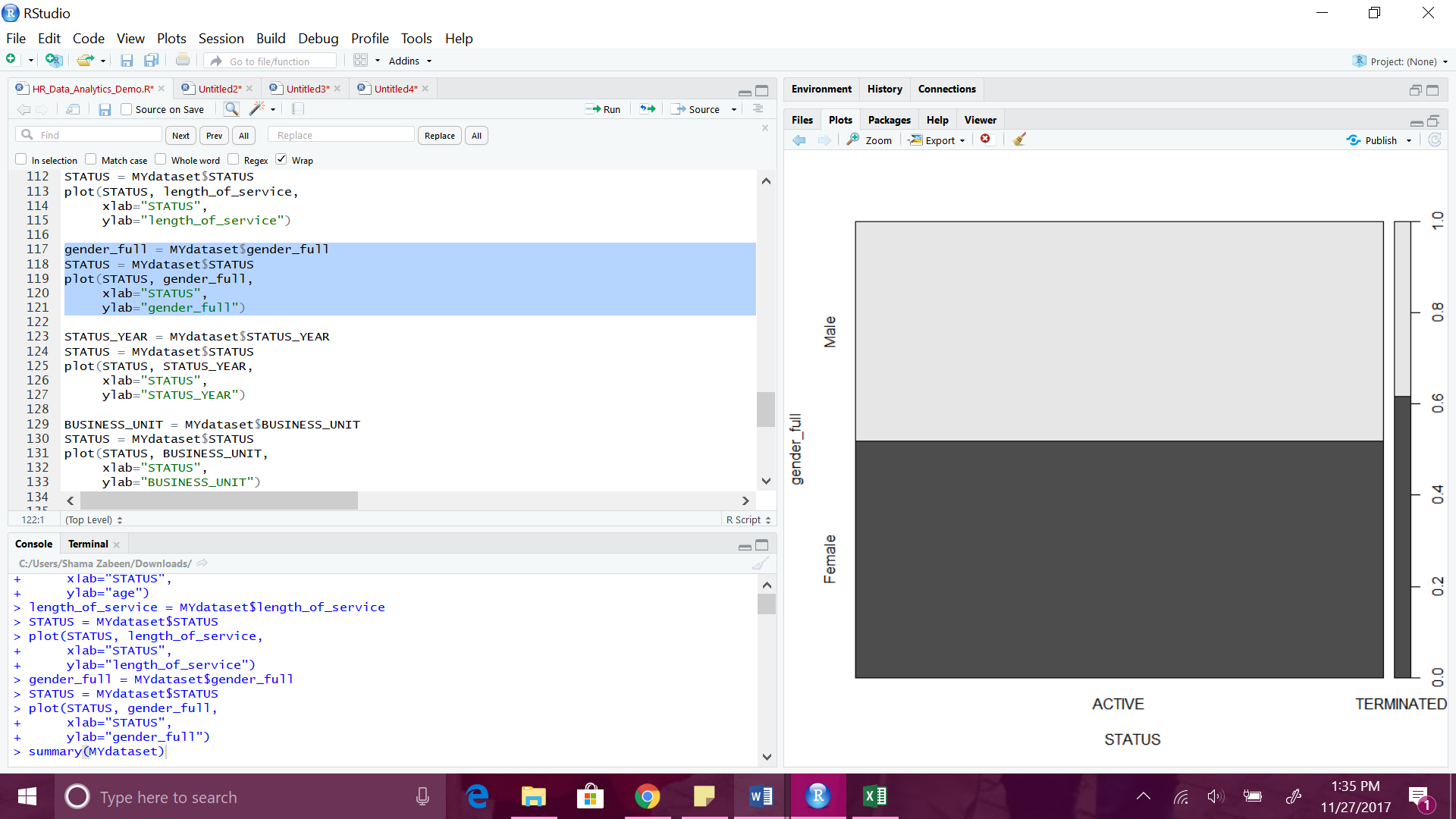
**STATUS = MYdataset$STATUS**

**plot(STATUS, gender\_full,**

**xlab="STATUS",**

**ylab="gender\_full")**

Figure 4.31 Based on Status and Gender

****

**TESTING THE TRAINED DATA BASED ON STATUS AND YEAR:**

**STATUS\_YEAR = MYdataset$STATUS\_YEAR**

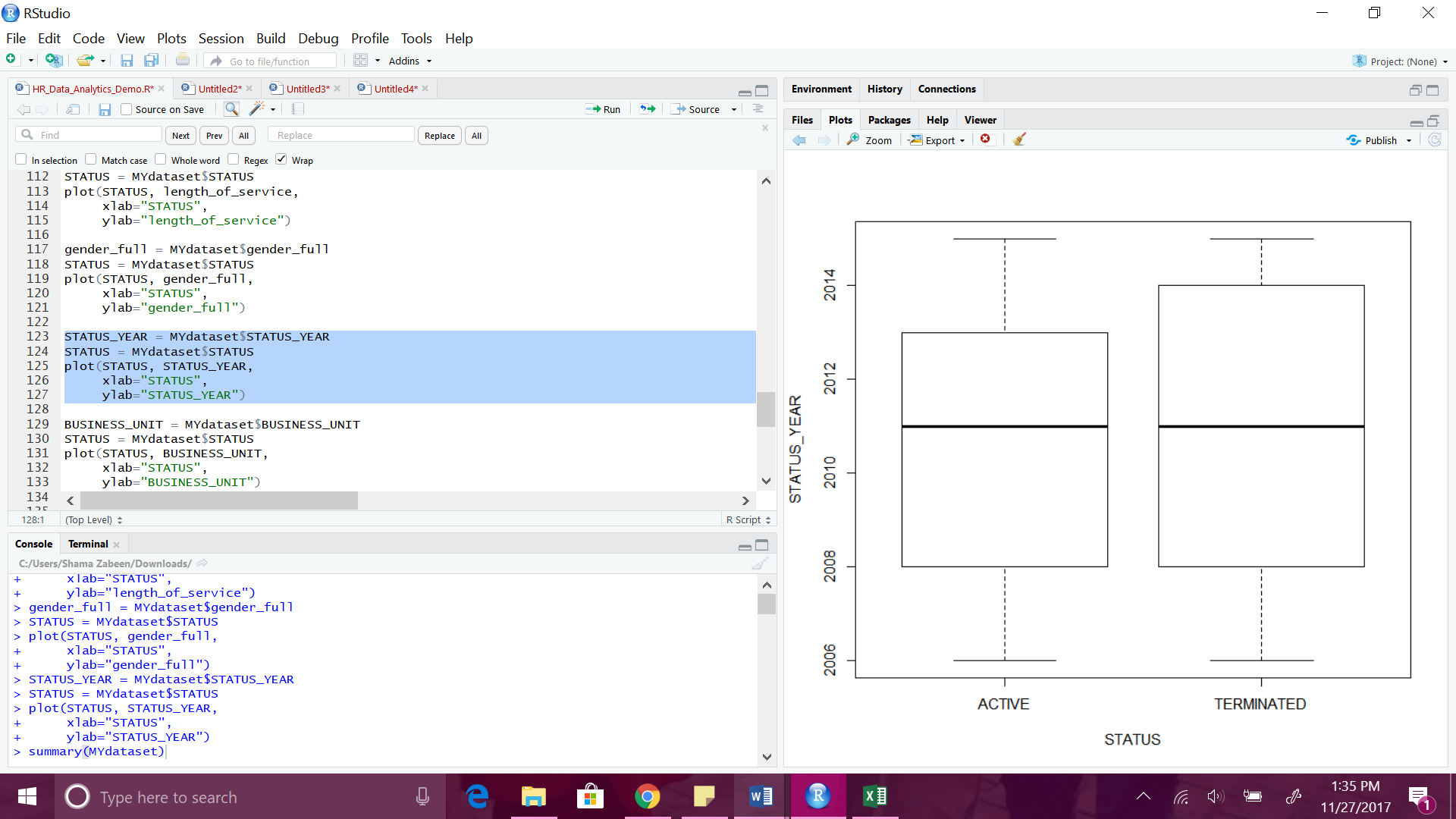
**STATUS = MYdataset$STATUS**

**plot(STATUS, STATUS\_YEAR,**

**xlab="STATUS",**

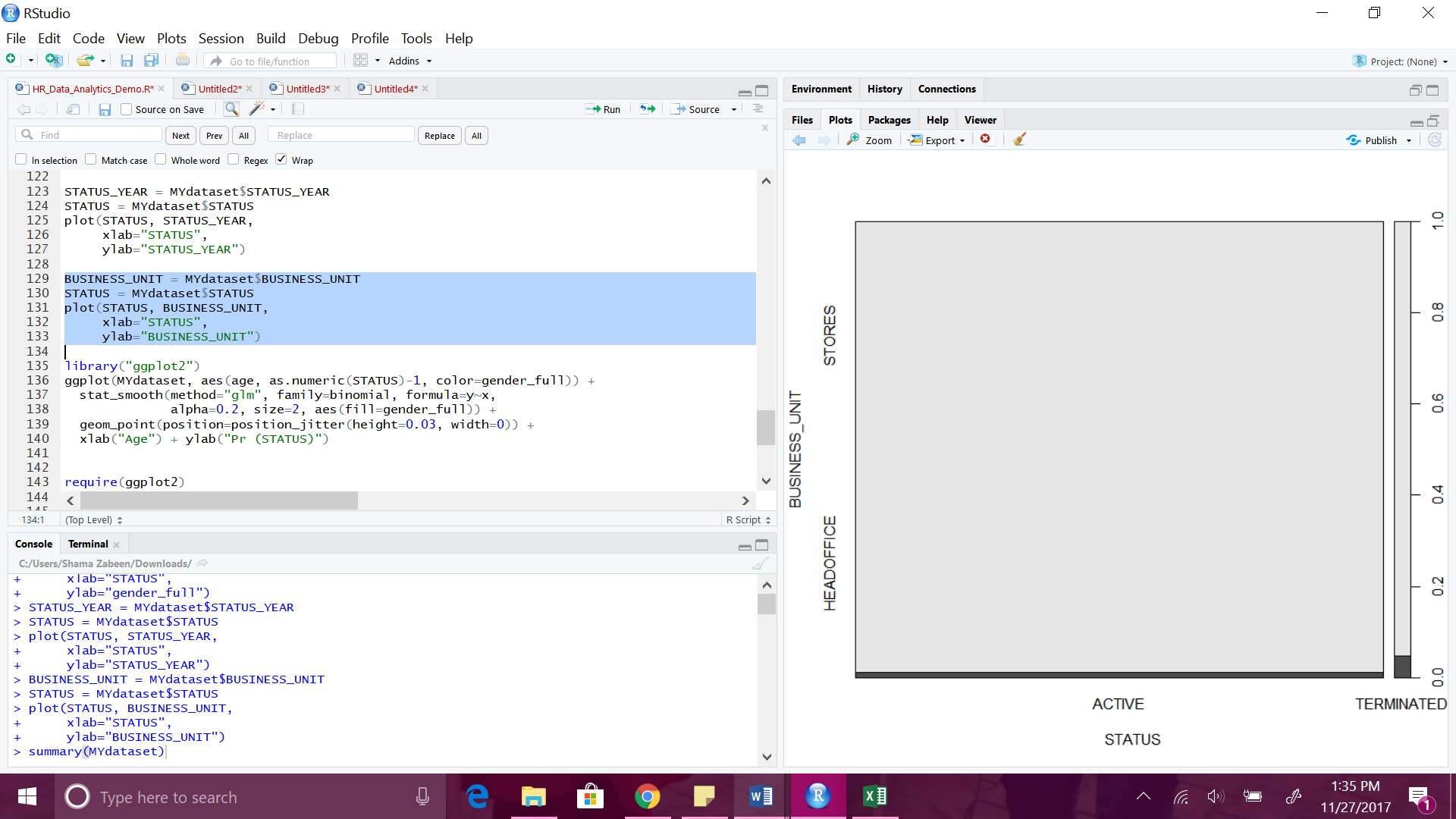
**ylab="STATUS\_YEAR")**

Figure 4.32 Based on Status and Year



**TESTING THE TRAINED DATA BASED ON STATUS AND LOCATION:**

Figure 4.33 Based on Status and Location

****

**TESTING THE TRAINED DATA BASED ON STATUS AND AGE AND LOCATION:**

**library("ggplot2")**

**ggplot(MYdataset, aes(age, as.numeric(STATUS)-1, color=gender\_full)) +**

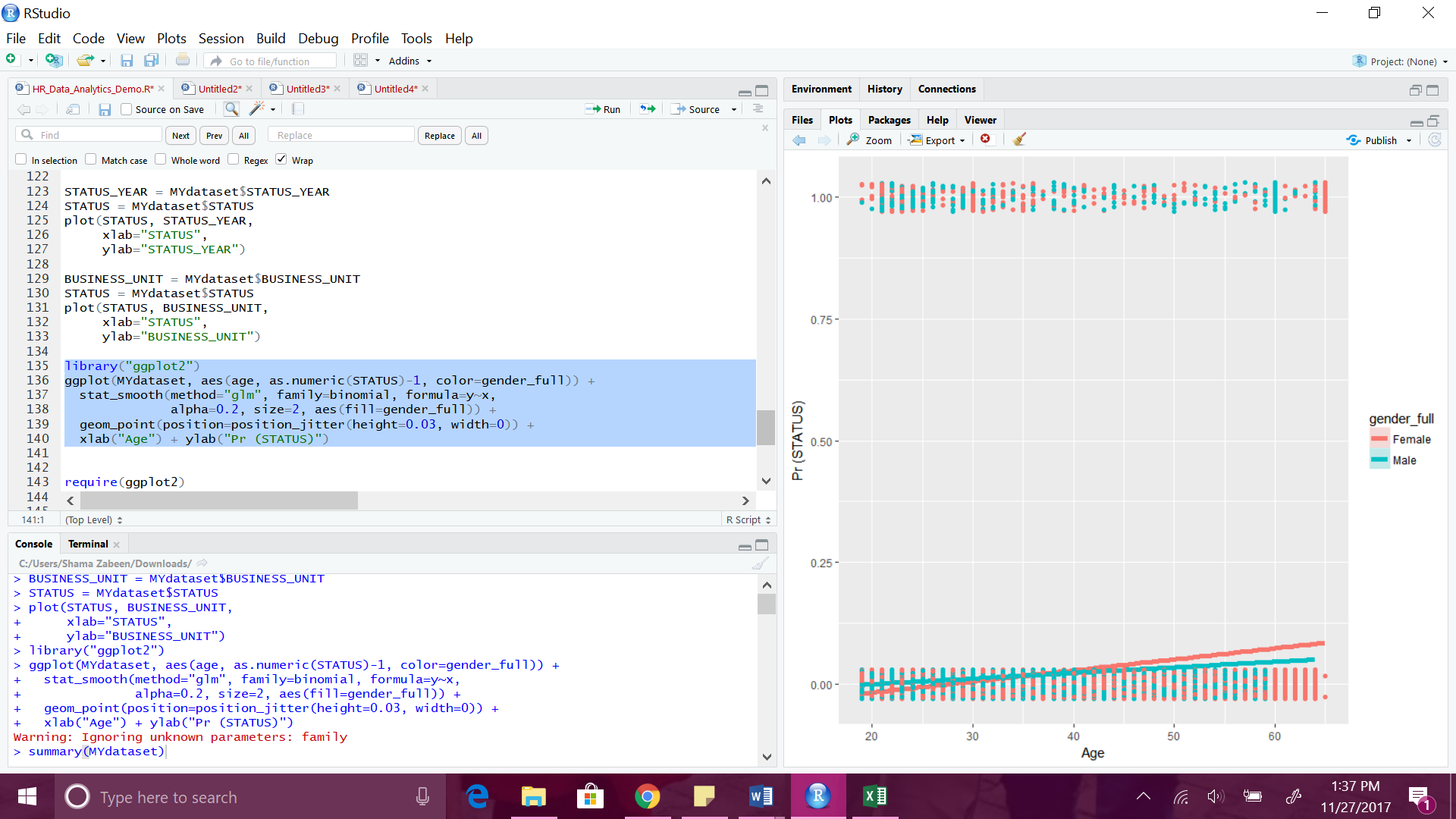
**stat\_smooth(method="glm", family=binomial, formula=y~x,**

**alpha=0.2, size=2, aes(fill=gender\_full)) +**

**geom\_point(position=position\_jitter(height=0.03, width=0)) +**

**xlab("Age") + ylab("Pr (STATUS)")**

Figure 4.34 Based on Status Age and Gender



**require(ggplot2)**

**ggplot(MYdataset, aes(age, as.numeric(STATUS)-1, color=gender\_full)) +**

**stat\_smooth(method="loess", formula=y~x,**

**alpha=0.2, size=2, aes(fill=gender\_full)) +**

**geom\_point(position=position\_jitter(height=0.03, width=0)) +**

**xlab("Age") + ylab("Pr (STATUS)")**

**ggplot(MYdataset, aes(x=age, y=STATUS))+geom\_point(size=2, alpha=0.4)+**

**stat\_smooth(method="loess", colour="blue", size=1.5)+**

**xlab("age")+**

**ylab("STATUS")+**

**theme\_bw()**

**qplot(x=age, y=STATUS, data = MYdataset)**

There were 93 predictions that were noticed for termination for the year 2015 based on the predicted data which was 985 accurate and the testing and the training were successful in the functions they were implemented.

**Data operations with python**

I used python mainly for five types of tasks in data operations:

1. Search for coordinates of cities to plot in the map box

For location related fields, for example, the Personnel Area, the single column was split first with Tableau to separate columns containing city, state, and country respectively.

Use geopy.geocoders package of python to search for addresses formed by “City, State, Country”. Latitude, longitude, and a more detailed address could be obtained. The results need to be double checked to ensure the correct city was exported.

Please refer to the GetCoordinates.py or the screenshot in Figure 1 for details. Figure 2 shows the dictionary generated. Some city names were spelled incorrectly, so need to be corrected to find the correct address.

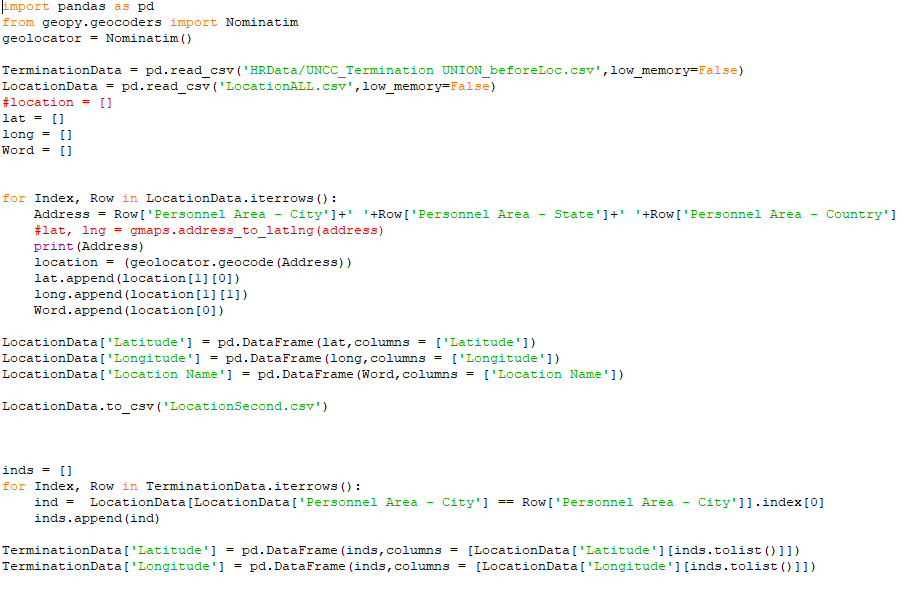


Figure 1, GetCoordinates.py

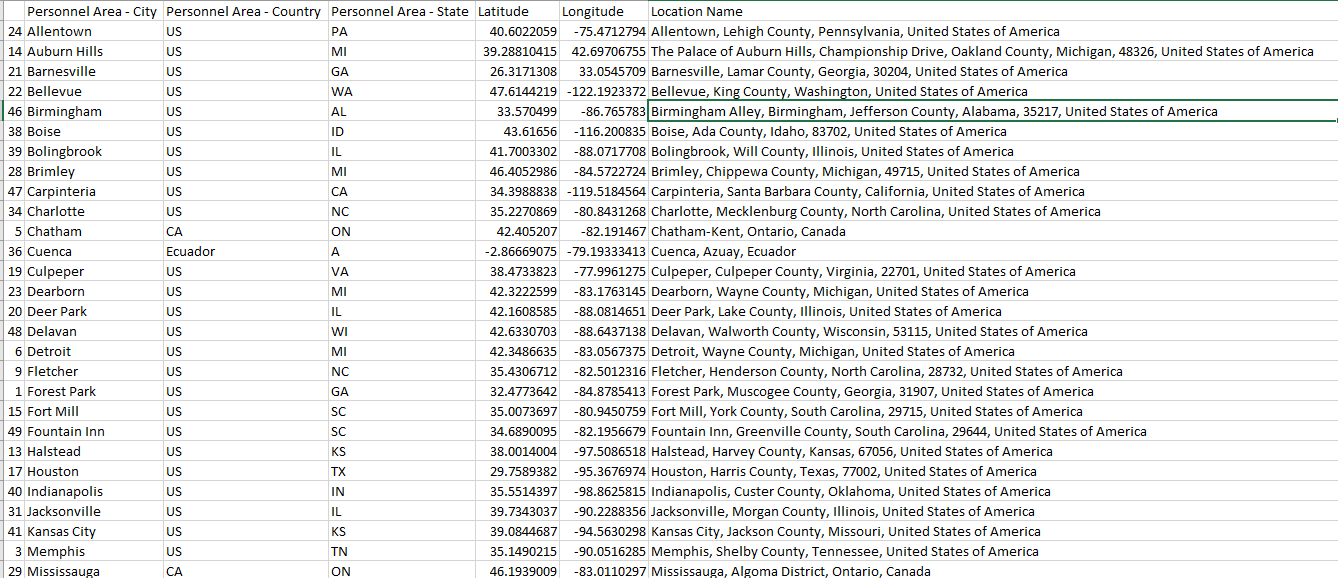


Figure 2, part of the address dictionary

1. Make dictionaries and fill in missing fields for a specific file or after merging files

The dataset contains a lot of code and corresponding names, for example, the Division code and Division, or the address book formed in 1. To fill in a file that contains the code but not the name or the details, a dictionary was made up containing a list of the code and the corresponding details to look up, for example, the address book in Figure 2. Then the field to fill up in a document is indexed through, corresponding code in the dictionary is located, and the new fields or empty fields were filled with the corresponding values in the dictionary.

Please refer to DictionarySearch.py or the screen shot in Figure 3.

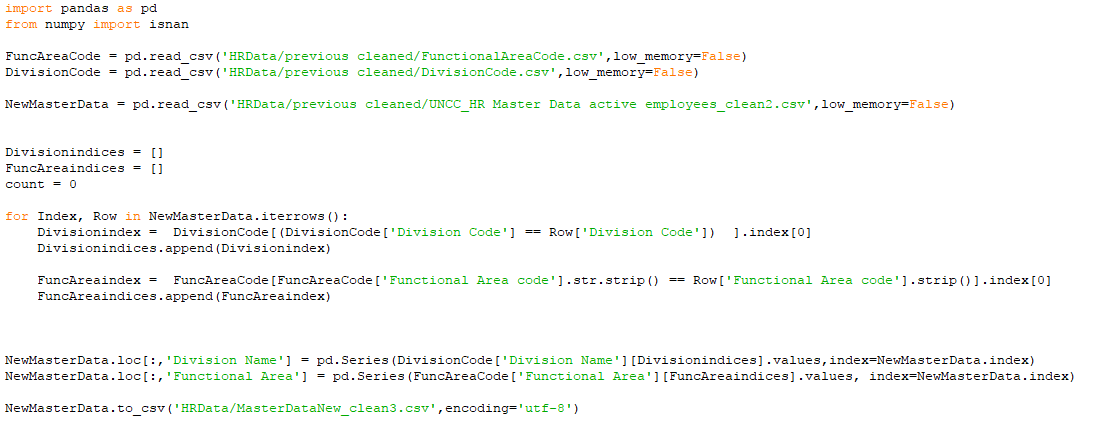


Figure 3, DictionarySearch.py

The termination data with addresses filled is partially presented in Figure 4, or please refer to the UNCC\_Termination Union\_loc.csv.

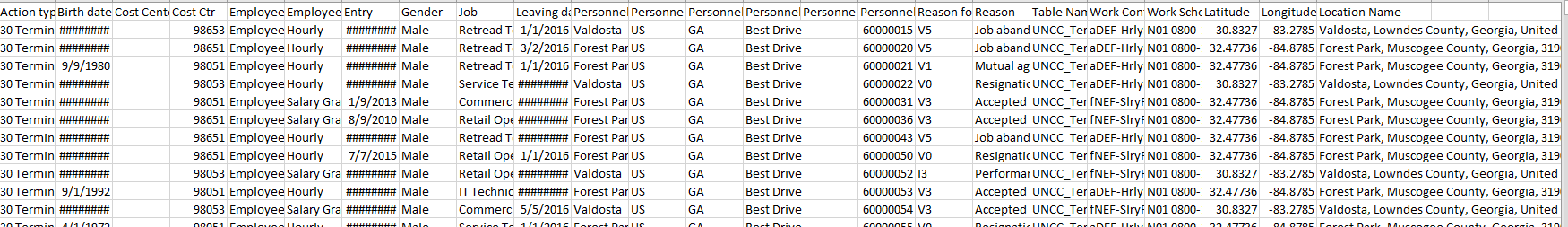


Figure 4, fill in the location details (Latitude and Longitude) in Termination data

1. Merge files

Merging file could be conducted in Tableau or python. For example, the absence data is merged with the master data to review more information of the absent employees. It is based on indexing the absence data based on personnel ID and find the corresponding record in the master data.

Please refer to AbsenceUnion.py or the screenshot in Figure 5.

Figure 5, merging absence data and master data.

1. Calculating specific fields and filtering

Python offers a lot of packages for various calculations. For example, the time inputs as a string could be transformed into time stamp formats and be used to calculate fields as age, length of time, or just extract the month or date. Null fields could be easily filled. Selected columns could be deleted. The dataset could be grouped and then filtered by different columns. In fact, all the operations in Excel and Tableau could be realized by coding with python so no data operations with other software are needed if well designed.

Please refer to the data loading and generating section in app.py or the screenshot in Figure 6 for some operation examples.



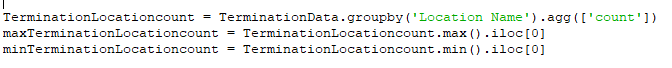


Figure 6, data operations with python

1. Predictive modeling

Python sklearn package offers powerful models and functions to do predictive modeling and analyses. Please refer to the make\_individual\_figure function in app.py or the screenshot in Figure 7 for an example of a prediction based on linear regression.



Figure 7, a predictive analysis based on linear regression

Interactive dashboard with python Dash

The major task accomplished with python is the dashboard. It is powered by a python framework Dash by plotly to build web applications. Please refer to this link for more details about Dash. <https://plot.ly/products/dash/>.

Two major aspects in the dashboard section are creating app components and creating callback functions.

1. Creating app components/layout

Components such as filter selectors, buttons, and graphs are defined in this section. Their locations, styles, and value set ups could be customized. For example, Figure 8 shows the code for defining the logo, title, and three tabs on the page.



Figure 8, app layout with a logo, title, three tabs, and the tab contents.

The components on each tab could be defined in the callback function of clicking each tab. The components used in this dashboard app includes range sliders, sliders, radio options, dropdown lists, buttons, texts, tables, and graphs. Figure 9 and 10 shows the examples of some components.



Figure 9, example of a filter set up with a combination of radio options and a dropdown list



Figure 10, example of the set up of two graphs in a row

1. Creating callback functions

Callback functions enable interactive responses of the app. Different components might be able to respond to different operations. For example, Figure 11 shows the callback function of the radio items in Figure 9.

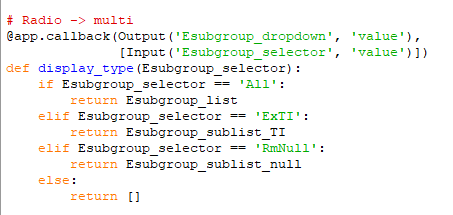


Figure 11, callback function of radio items

The graph callbacks could be more complicated and take in more inputs since they respond to all filters and contains the graph definition and layouts. Please refer to app.py or the screenshots in Figure 12 for an example of a graph updating according to multiple filers.



Figure 12, the map box graph to displaying number of termination at different locations

Deploy app with Heroku

The app was deployed and ran on Heroku for external access. Please refer to the following link for more details about deploying Dash with Heroku: https://plot.ly/dash/deployment.

The product: HR Predictive Modeling Dashboard

Figure 1 shows the current deployment of the demo and some labels explaining the functions.



Please refer to <https://my-dash-hrpred2.herokuapp.com/> to find out more functions and test the features.

**SUMMARY:**

The testing of the dataset was done to an accuracy of 98.4% which states that the training algorithms were very correctly implemented. Also, the main feature of this language R that was very useful was the inbuilt library functions. The usage of these inbuilt functions has made the length of the code way too lesser and the time complexity and space complexity were minimized compared to other programming languages that we have earlier worked on.

As we have here made use of the decision tree model, with respect to the future scope, Random forest, Boosted Models, Support Vector Models and Linear Models could be implemented on the same data set to get better results.