IT-Helpdesk Resolution Time Analysis

Presented By:

Arundhathi Patil
Chinmayi Amit Karmalkar
Manisha Nln
Manoj Angane
Swayam Joshi

9th Nov 2018

Dr. Anol Bhattacherjee

Executive Summary

Completing work on time is everyone's aim. Yet, sometimes due to hardware or software issues things gets delayed, thus causing pushbacks in work. With the world progressing towards mechanizing every aspect of daily lives, we are attempting here to get a glimpse of how this technique is going so far in terms of Helpdesk in the IT industry.

Collection of data and data source play a paramount role in the analysis since analysis relies on data and data has been to be legitimate, else the entire analysis would be a waste. To prevent such an event in this report, we have taken the data set from IBM Watson.

With analysis of IT Helpdesk data, we aim to understand the cause and relationship between various problems faced by the IT industry, especially when it comes to resolving issues and closing tickets. We also aim to provide worthy recommendations to the IT Helpdesk, to encourage them to improve in their work, thus caring about every other technology using a person's time and patience. The data is a register/log-book of tickets raised with specific details about each ticket and the employee who raised it. We performed exploratory analysis, built hypothesis and regression models to analyze the data. The data modelling involved building up of several models with interactions between independent variables to find out the best model to provide feasible recommendations for better response time for all employees.

The key findings from our analysis suggest that the response time varies with the functional domain that ticket belongs to and has a significant impact of seniority of the employee raising the ticket.

Table of Contents

PROBLEM SIGNIFICANCE

The IT industry booms with the increase in quarterly gains and publish the results. Yet what runs these companies are the IT helpdesk and their resolution of issues faster and effectively. To get this clarity in more depth and understanding, we have tried to figure out some problems and come up with a solution.

Having worked in the IT industry, it is acknowledged how much working hardware and software is essential. This brings us to the problem we aim in understanding well and hopefully provide solutions on this set.

In every IT help desk, there are always new issues that need fixing, some that are pending and some that might need additional data. Considering these, we have some insight about the complaint data, such as ticket type, severity, filed against and satisfaction after resolution.

The analysis is aimed at understanding where the maximum time is consumed when an issue gets reported to the helpdesk and to analyze how to reduce the number of days the complaint was open. We also want to see how different factors affect the solvability of the issue and if there are any other factors that might affect the data analysis.

With the provision of ticket/service management tools and multiple communication services ranging from remote desktop support to instant messaging applications, the IT helpdesk is expected to perform much better. But is technology and smooth communication service the only means to improve response time is a question left to ponder about. While we face challenges of omitted variables bias in our data about the time of the day and day of the week the issue was raised and sentiment of the requestor along the course of time the issue was resolved, the reports focuses on the basic indicators and factors that could help increase the efficiency of helpdesk and save capital by enabling the workforce to deliver on time.

DATA SOURCE / PREPARATION

The data source we have considered is taken from IBM Watson. It was collected in the year 2015 for analyzing how fast/slow the issues get resolved and measure the satisfaction index of it. The data comprises of 100,000 closed tickets that were filed at their help desk. The data provided is relatively clean and does not require cleaning excessively. Although it could have had more columns included in the data set to allow better modelling, this data set helps for the preliminary understanding nevertheless.

The data missing from this dataset includes the time of putting in the request and the time when the request was given attention to. This would have helped analyze aspects such as time to resolution and might have led in some direction towards a better solution provided.

The data present in the Severity, Priority, Satisfaction and Requestor Seniority was available in words with each word giving a degree/measure of the variable with respect to each ticket. We converted these variables into levels for better processing of data. The dependent variable was heavily skewed, so we applied a transformation technique to handle the skewness. There were no outliers in our data after the transformation, so outlier handling was not an issue and there were no missing values, invalid data for any variable to be cleaned or imputed.

The variables used include some of the following mentioned variables:

Column Name	Description
daysOpen (Dependent Variable)	***The number if days the ticket was open before being resolved or closed***
Requestor	*Employee ID who submitted the ticket*
ITOwner	*Employee ID of IT employee who serviced ticket*
FiledAgainst	*Functional area the ticket was filed*
Severity	*Submitter assigned severity of ticket the assignment of value (from 0 to 4) is as follows* 0 - unclassified 1-minor 2-Normal 3- Major 4 - Critical
Priority	*IT assigned priority of ticket, this is the variable that the IT helpdesk can control to improve time taken to resolve an issue. The assignment of value (from 0-3) is as follows:* 0 - unclassified 1-low 2-medium 3- high
Satisfaction	*It is the satisfaction level of the employee who raises the ticket* 0 - unknown 1-unsatisfied 2-satisfied 3- highly satisfied
Requestor Seniority	*The Seniority of the employee who raises the ticket is assigned as:* 1 - junior 2 - regular 3 - senior 4 - management

Note:
The Requestor and ITOwner variables have not been considered for further analysis as they would be more helpful to analyse individual performance which might narrow down the scope of this report.

HYPOTHESIS:

In this section we are going to discuss some of the factors that might influence the target variable - "DaysOpen" - the number of days it took for the IT Helpdesk to resolve the issue and whether these factors have a positive or a negative effect on the target variable.

We are interested in focusing on the actionable variables while forming the hypothesis. The actionable variables are the variables on which the IT Helpdesk have a control over and based on the results the IT Helpdesk can make necessary changes in the management and resolution of the tickets. We are also considering some control variables which might have a significant effect on the target variable – "DaysOpen".

Based on our logic and experience in the IT industry, we could logically figure out that the priority assigned by the helpdesk pays a major role in how the resolution is carried forward. Moreover, depending on the functional area of the ticket, an automatic priority must be allotted for better and quick resolution. The severity assigned by the requestor might be impacting the priority assigned to the ticket as well.

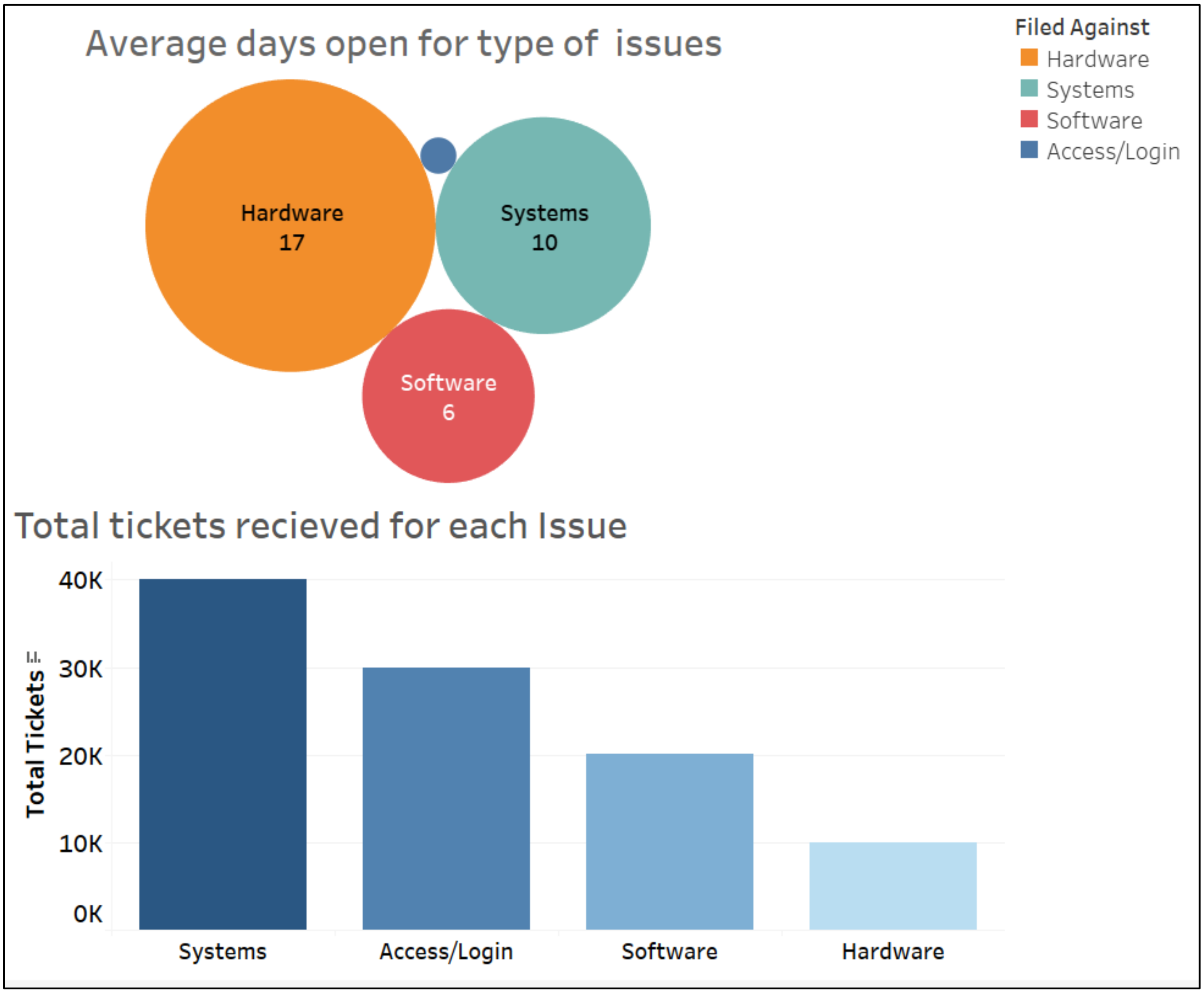
The following hypothesis has been considered to be tested for this project:

Hypothesis	Description
1. H1a: βPriority < 0	The priority of a ticket is set by the IT Helpdesk. It ranges from values 0 to 3 with 3 being the highest priority ticket. With an increase in priority for a ticket, it is expected that the ticket should be resolved sooner. Hence DaysOpen value should reduce with an increase in priority.
2. H2a: βFiledAgainst ≠ 0	FiledAgainst describes the issue for which this ticket is raised. We want to test if the number of days the IT Helpdesk team takes to resolve an issue/request varies with the functional area the ticket was filed for.
3. H3a: βPriority * Severity < 0	For high severity tickets priorities should be also set high. It is expected that a ticket should be resolved quickly with an increase in Priority and Severity of the ticket. Hence the DaysOpen value should reduce in such scenario.

Note:
RequestorSeniority variable has been considered as a control variable to check if there is an influence of Seniority of the person who has filed the ticket.

DESCRIPTIVE ANALYSIS

In the initial exploratory analysis, we found the following distribution for total tickets received.

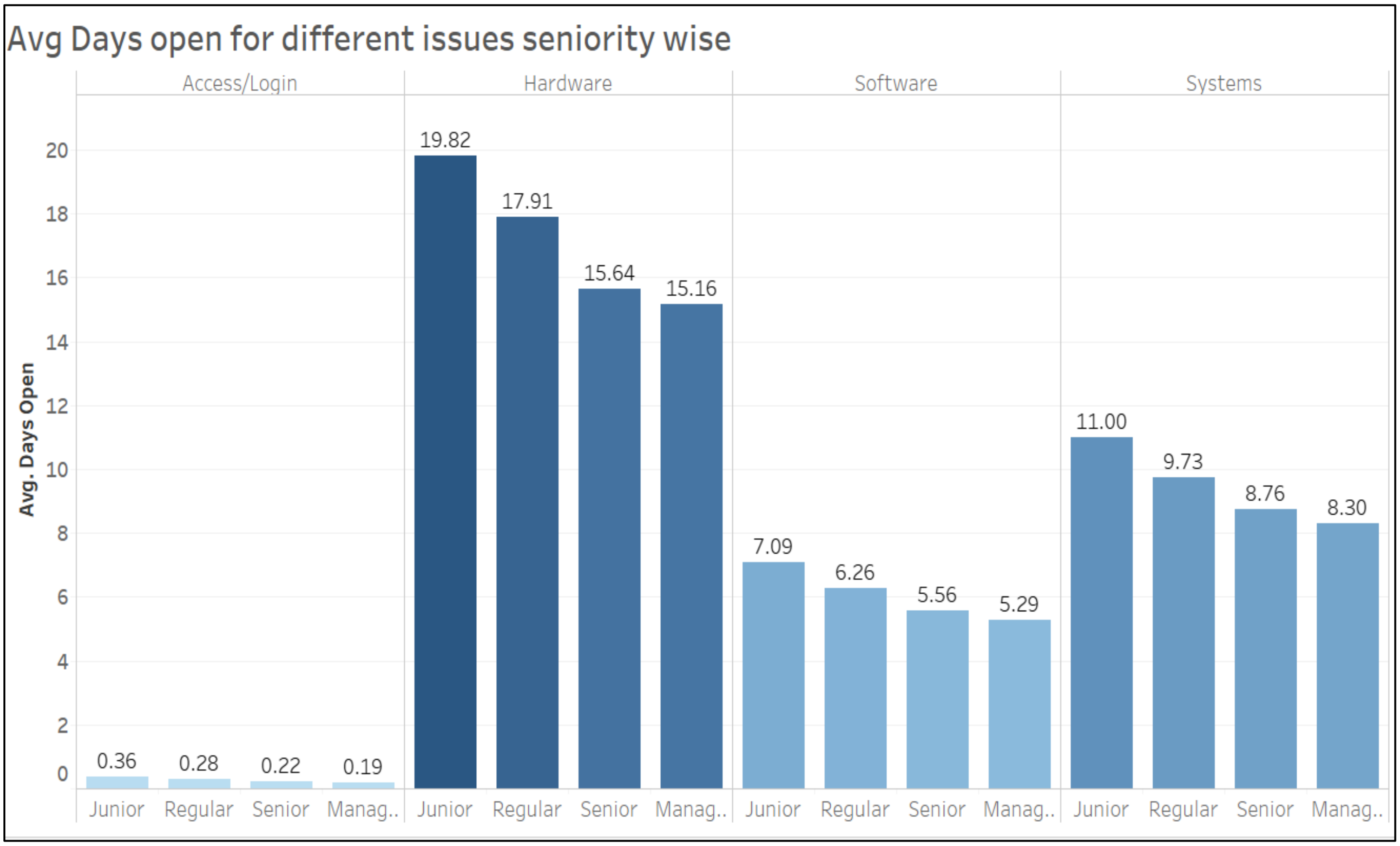


- The above visualizations show the distribution of tickets based on the Functional area of the ticket. The issues related to Hardware have been open for a longer time which implies that IT Helpdesk took a comparatively longer time than usual to resolve Hardware issues. The IT Helpdesk team took almost the same average number of days to resolve issues related to Systems and Software. The access and login issues took very less time to resolve.

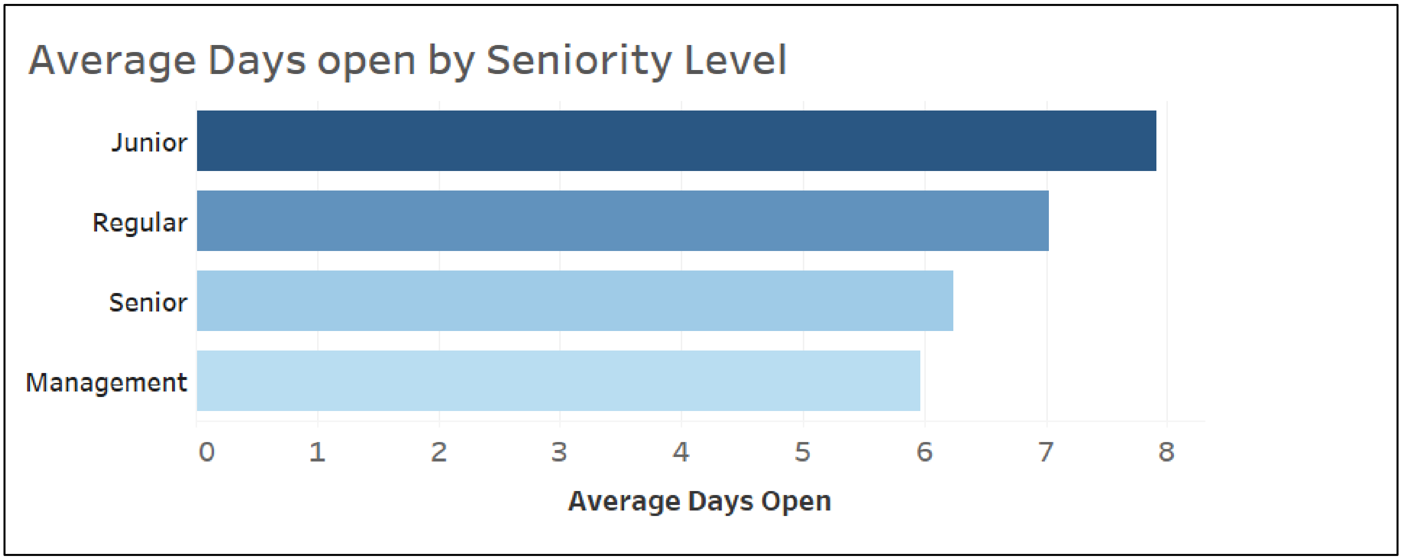
- The bar graph shows that Systems issues Category has the highest number of tickets followed by Access/Login and Software Issues Category.

- The hardware issues are comparatively very less, indicating good hardware procurement, but the tech support for hardware is still under question looking at the plots above.

Since we are considering Seniority as a factor for assessing the resolution of tickets raised, we found some interesting insights:



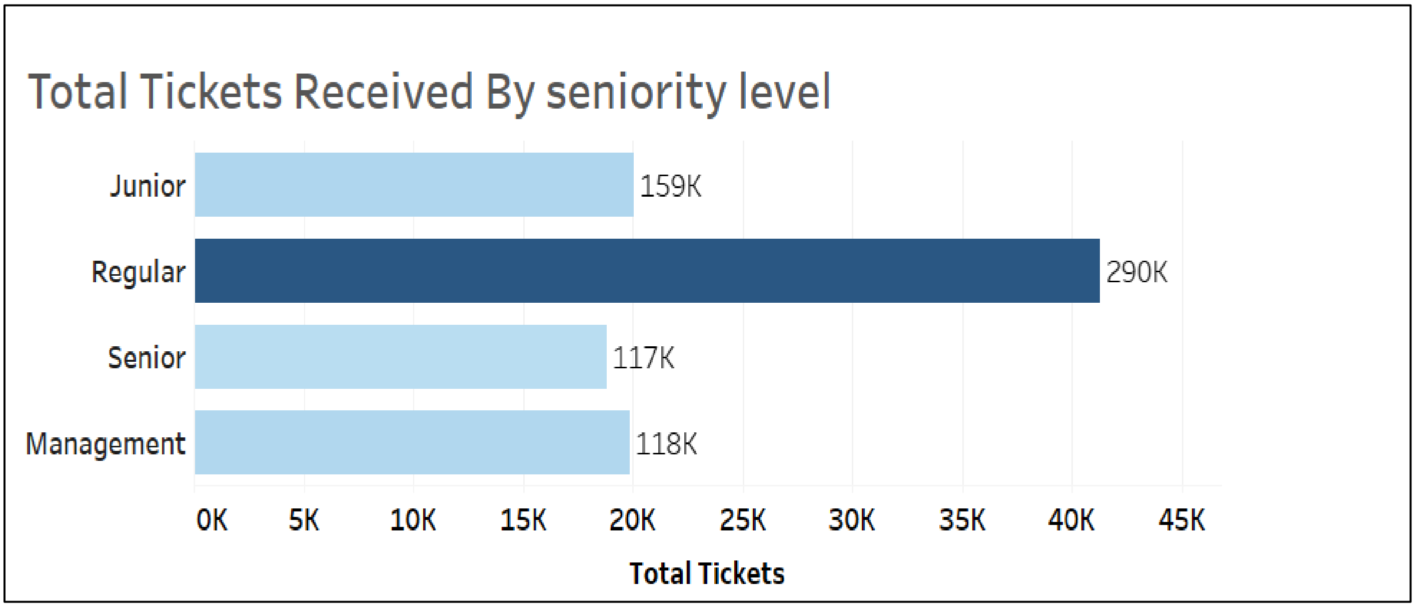
- The graphs clearly show that tickets filed by a Junior level employee have been taken maximum time to be resolved by the IT Helpdesk. This trend can be seen consistently for tickets related to all types of issues like Hardware, Software, Systems and Access/Login. It seems there is some influence of Seniority on the time taken to resolve the tickets by IT Helpdesk based on this graph.



- The graphs clearly show that tickets filed by a Junior level employee have been taken maximum time to be resolved by the IT Helpdesk. This trend can be seen consistently for tickets related to all types of issues like Hardware, Software, Systems and Access/Login. It seems there is some influence of Seniority on the time taken to resolve the tickets by IT Helpdesk based on this graph.

The data we are analyzing has many independent variables of which we have the correlation matrix for the following:

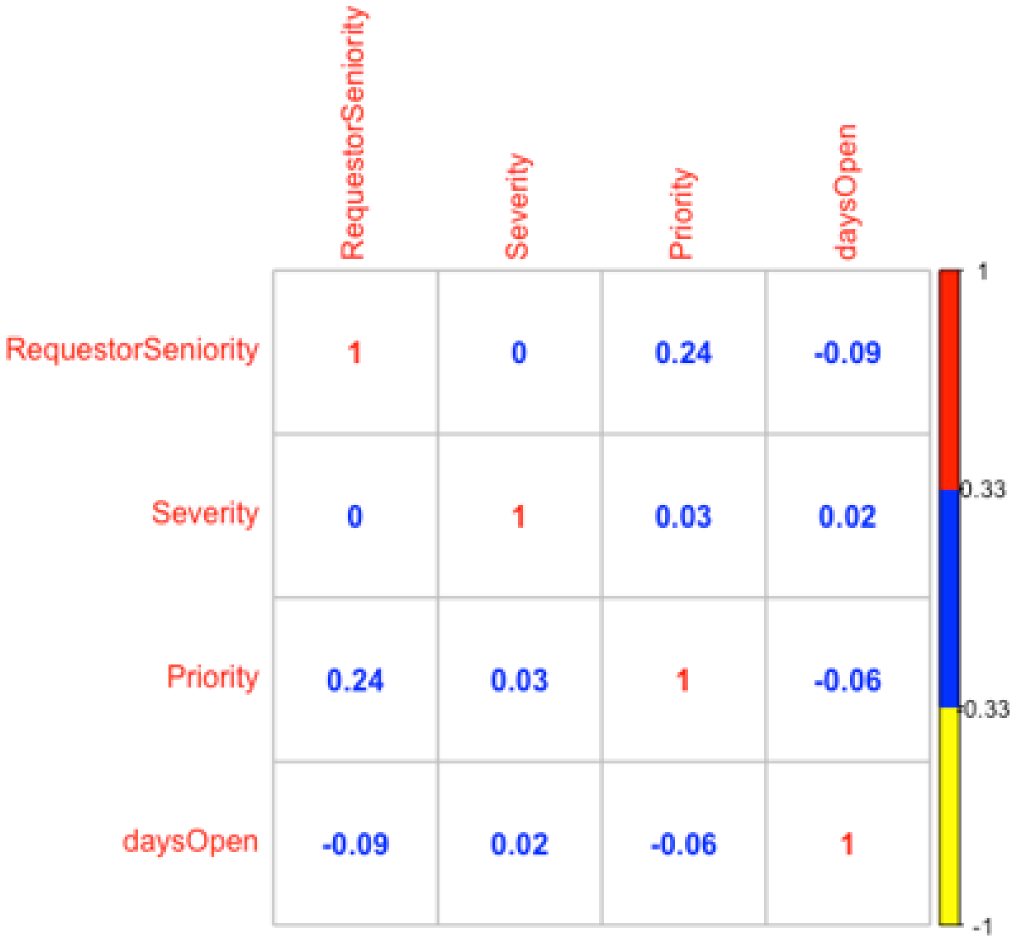
- Requestor Seniority
- Severity
- Priority
- Days Open



- The highest number of tickets have been filed by Regular level employees. This shows that a lot of IT issues are being faced by Regular level employees.

- On our analysis so far, the number of tickets raised by the major workforce which is the regular employee category take considerably a higher time to resolve impacting the overall efficiency of the helpdesk service and indicating discrimination in resolving a ticket based on seniority

Plotting a correlation matrix for the given set of variables under consideration, we get something as follows:



	RequestorSeniority	Severity	Priority	daysOpen
RequestorSeniority	1	0	0.24	-0.09
Severity	0	1	0.03	0.02
Priority	0.24	0.03	1	-0.06
daysOpen	-0.09	0.02	-0.06	1

- The correlation terms are not high for any term that might suggest multicollinearity.
- The correlation between the terms is positive for Requestor Seniority and Priority, for Severity and Priority, Severity and daysOpen, which mean they are directly related.
- When we check for correlation between terms like Requestor Seniority and daysOpen and daysOpen with Priority, we observe negative correlation. This suggests that we have indirect relationship between these terms and that they are inversely related with one another.
- From the correlation matrix given we can say that there is no issue of multicollinearity in the data for predictor variables. This helps us cognize that the independent variables cannot be predicted by other variables.

variables.

MODELS:

MODEL BUILDING

Based on the above hypotheses and the descriptive analysis, the following model has been formulated:

Since the target variable is continuous variable providing information about days taken to close IT helpdesk data and the model has been built using Regression and the analysis has been done in R.
When we studied our variable (daysOpen), its distribution shows exponential decreasing curve and so preferred using transformation for making distribution normal.
After our analysis of distribution curve, we used logarithmic transformation making distribution somewhat normal.

As our data contained few values for daysOpen as 0, for smooth transformation we added “1” to daysOpen before transforming data. So, our model will predict log(daysOpen+1) as dependent Variable.

MODEL 1

Regression model predicting log(daysOpen+1) based on priority, RequestorSeniority, Priority interacting with Severity, FiledAgainst and TicketType.

MODEL 2

Generalized Regression model predicting log(daysOpen+1) based on priority, RequestorSeniority, Priority interacting with Severity, FiledAgainst and TicketType with weights for making residuals follow Homoscedasticity.
We used generalized model to alleviate effect not 100% normal distribution as this model can eases this assumption.
We applied weights to reduce spread off residual plot and tried to make model follow assumption of normal and random distribution of residuals.

MODEL COMPARISION

The two models built to perform a regression analysis: Linear Regression and a GLM model with Weights

The measure of goodness we used to compare the models are

AIC Value: The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. It is a measure of how parsimonious the model is relative to the other models used to model the same data set. The lower the AIC value, better the model is.

BIC Value: Bayesian information criterion (BIC) is a similar criterion to select a model from a given set of models used to perform regression over a given dataset. Like AIC, lower the BIC model, better the model is.

Both the criterion penalize a model for adding more parameters to fit the data. As adding more parameters leads to overfitting.

The Application of weights is an attempt to make the model homoscedastic.

Comparison

Model	AIC	BIC	R-Squared
Linear Regression	138983	139078.1	**80%**
GLM Model (With weights)	123382	123475.9	Not Applicable

As we can clearly see, even though the R-squared value for the linear regression model is about 80%, the AIC and BIC value of the GLM model are lower rendering that the GLM model is relatively better than the linear Regression model.

QUALITY CHECKS

Independence check

As each ticker is an independent ticket as the resolution of one ticket dose not depend on the resolution of another ticker, the independence assumption has been satisfied.

Multi-collinearity

We performed tests, and all the variables considered in the model are categorical and are independent of each other, the assumption of multi-collinearity is satisfied.

Auto-correlation

All the analysis has been performed on the data in the year 2015 and hence there is no chance of Autocorrelation in this analysis.

Fisher Scoring:
The iteration value of 2 indicates that model had to iterate twice to successfully converge.

RECOMMENDATIONS:

1. High priority tickets are resolved in less time compared to low priority (β=0.008391) which justify our hypothesis.
2. The tickets with hardware and systems issues are taking more days to be resolved by the helpdesk in comparison to other issues. The helpdesk team should try to investigate as to why this is happening and take corrective actions to provide better service to the users for these types of tickets.
3. Priority and severity together are helping ticket to resolve quickly in accordance of the grievance of the issue.

OBSERVATIONS:

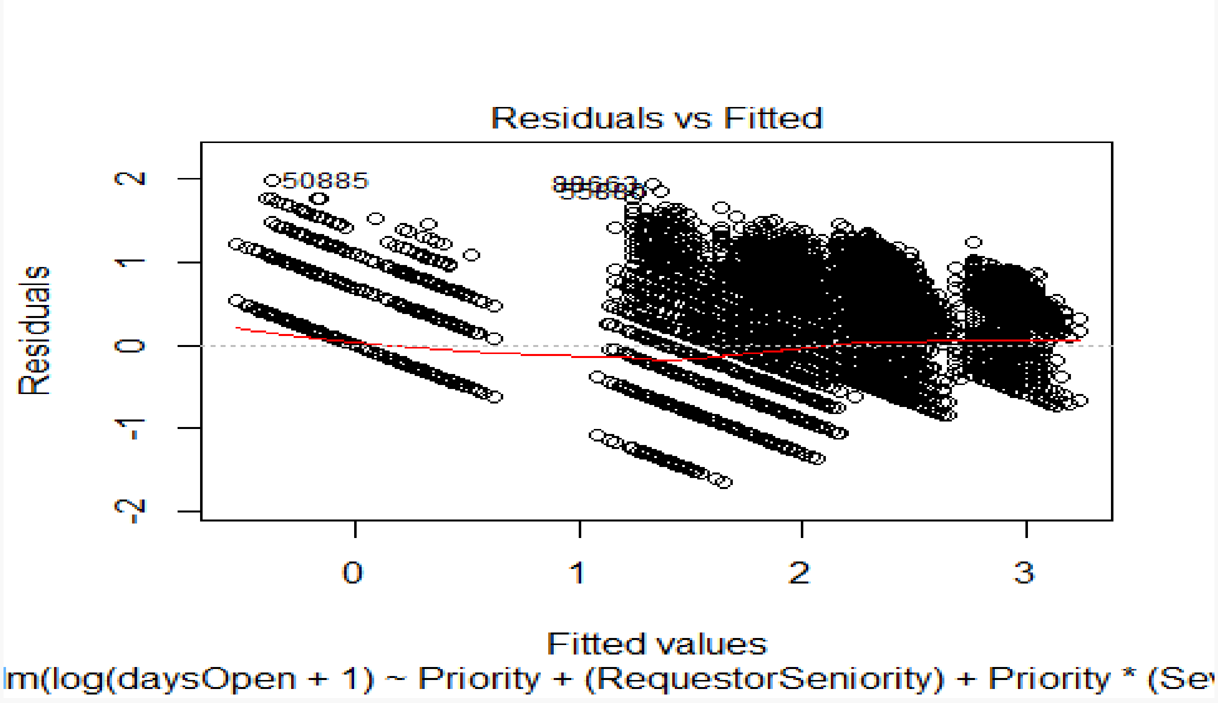
Based on our analysis, it shows that tickets raised by senior-level employees have been solved quickly as compared to other employees. This may be coincidence, but it shows significant impact on response variable daysOpen (-0.065047).

This significantly highlighted our observation.

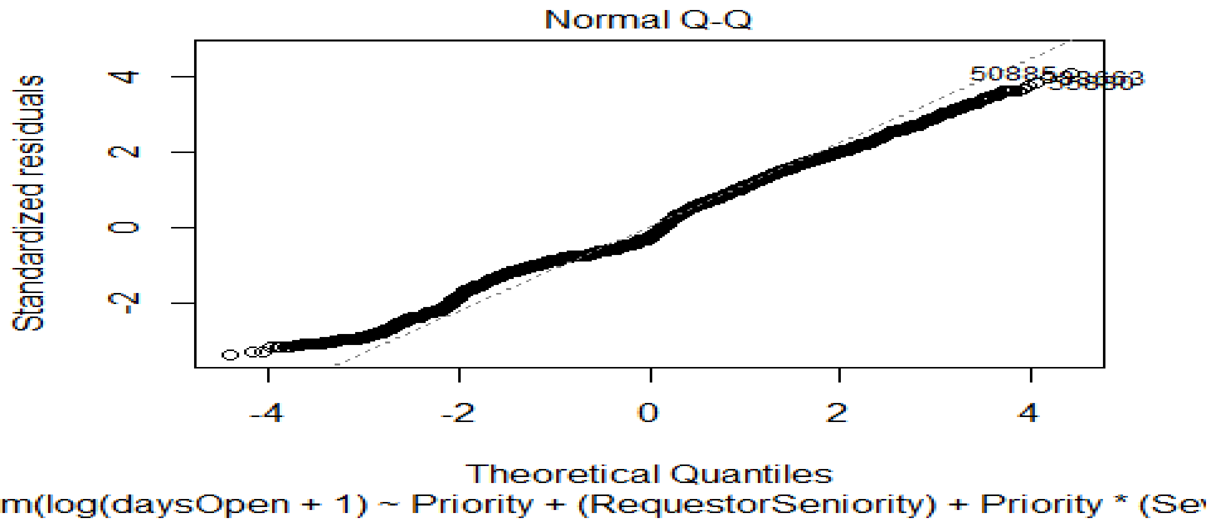
APPENDIX

Final Project Model

```
setwd("D:/Users/Swayam/CourseWrok/First Semester/SDM/Data Sets")

library(readxl)
it <- read_excel("IT_HelpDesk.xlsx")

m8 <- lm(log(daysOpen+1) ~Priority + (RequestorSeniority) + Priority*(Severity) + as.factor(FiledAgainst)+ as.factor(TicketType), data = it)
AIC(m8)

## [1] 138983
```



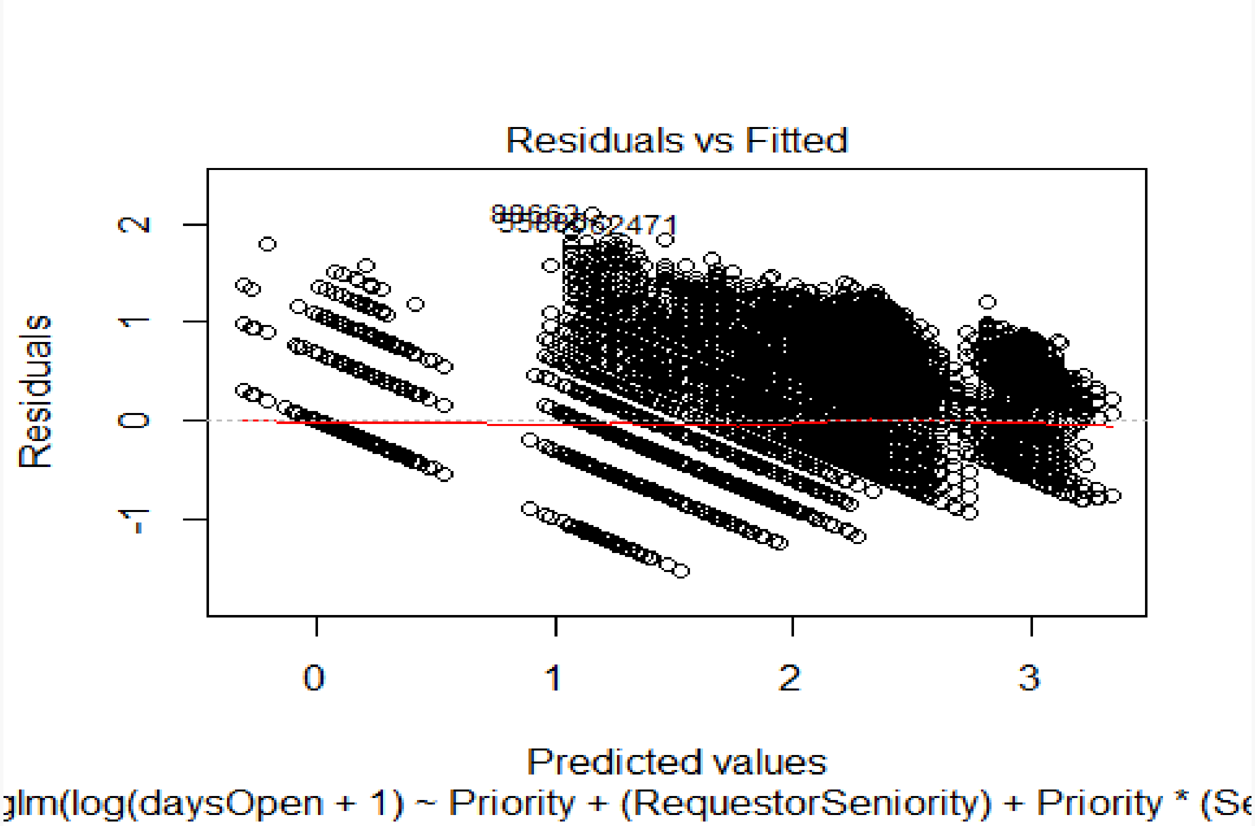
```
plot(m8)
```

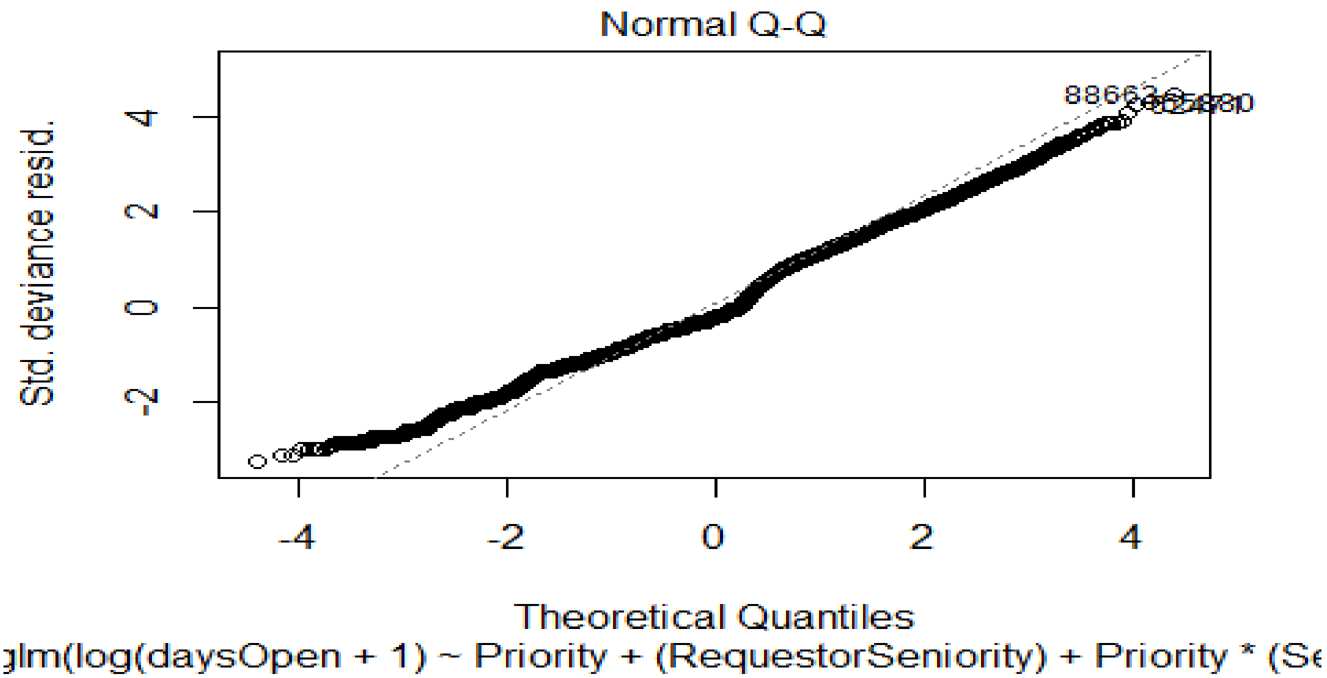


```
model <- glm(log(daysOpen+1) ~Priority + (RequestorSeniority) + Priority*(Severity) + as.factor(FiledAgainst)+ as.factor(
TicketType), data = it, family = gaussian, weights = m8$fitted.values^0.01)

summary(model)

##
## Call:
## glm(formula = log(daysOpen + 1) ~ Priority + (RequestorSeniority) +
##     Priority * (Severity) + as.factor(FiledAgainst) + as.factor(TicketType),
##     family = gaussian, data = it, weights = m8$fitted.values^0.01)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5285  -0.3264  -0.1073   0.3883   2.0990
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.615284   0.015473 -39.766  < 2e-16 ***
## Priority                    -0.008391   0.006953  -1.207  0.22746
## RequestorSeniority          -0.065047   0.001568 -41.486  < 2e-16 ***
## Severity                     0.119236   0.006846  17.418  < 2e-16 ***
## as.factor(FiledAgainst)Hardware  2.790316   0.005745 485.695  < 2e-16 ***
## as.factor(FiledAgainst)Software  1.793180   0.004695 381.913  < 2e-16 ***
## as.factor(FiledAgainst)Systems   2.190281   0.004064 538.907  < 2e-16 ***
## as.factor(TicketType)Request     0.745080   0.004097 181.838  < 2e-16 ***
## Priority:Severity           -0.010596   0.003323  -3.188  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2223516)
##
##     Null deviance: 101768  on 92633  degrees of freedom
## Residual deviance:  20595  on 92625  degrees of freedom
##   (7366 observations deleted due to missingness)
## AIC: 123382
##
## Number of Fisher Scoring iterations: 2
```

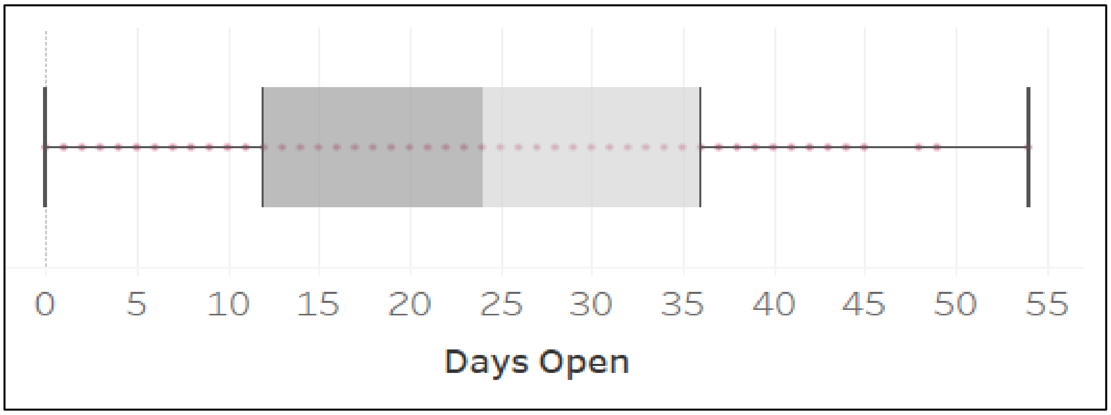
```
AIC(model)
```

```
## [1] 123381.5
plot(model)
```

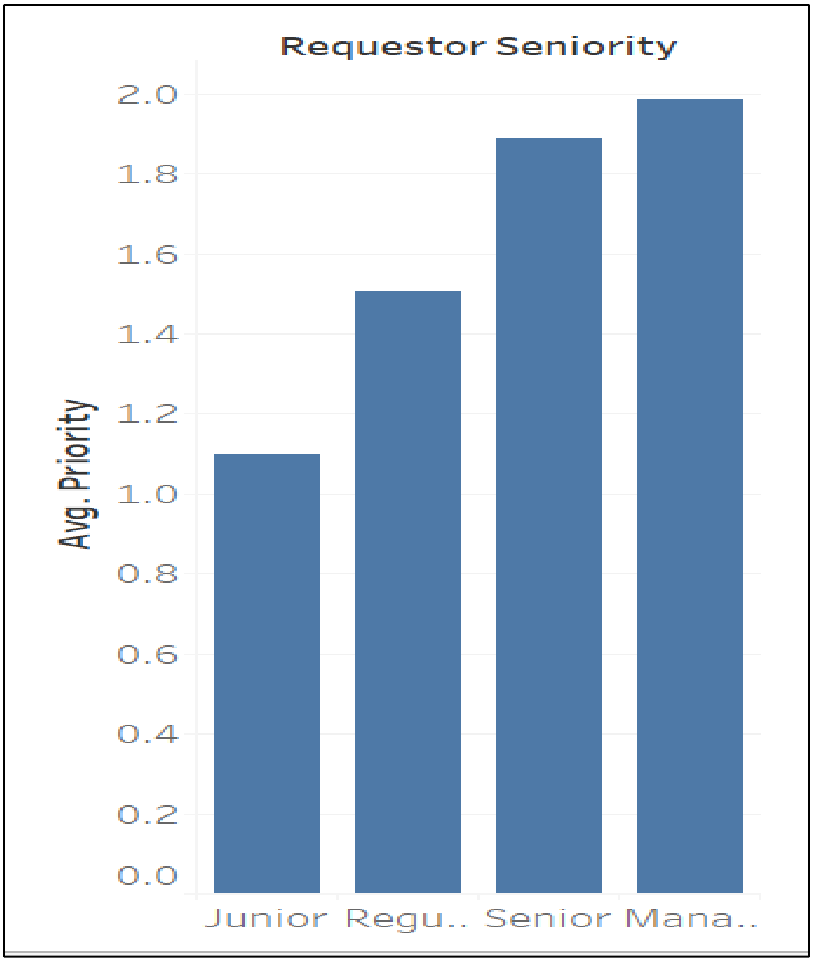




Outlier Analysis for Log(DaysOpen+1)



Priority Assigned as per Seniority


Requestor Seniority
Avg. Priority
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Junior
Regu..
Senior
Mana..