

**ASSIGNMENT FRONT SHEET**

**Course Name: ALY6000 72069 Introduction to Analytics**

**Professor Name: Dr.Prabagaran Santhanakrishnan**

**Student Name: Dong Quoc Tuong (Lukas)**

**Student Class: Fall 2019 CPS Term: A. 2019**

|  |
| --- |
| **Module 4: Careers in Analytics**  **Completion Date: October 13th Word Count: 1407 Due Time:12:00am** |

**Statement of Authorship**

*I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University’s regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.*

**Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

1. **Introduction**

The report is a practice to draft compelling data visualization of a set of data in R language. It is divided into three different parts and the data is comprised from two excel files: GWP\_Audits\_Data and CSQA\_Audits\_Data. It also provides business questions to for an auditing company.

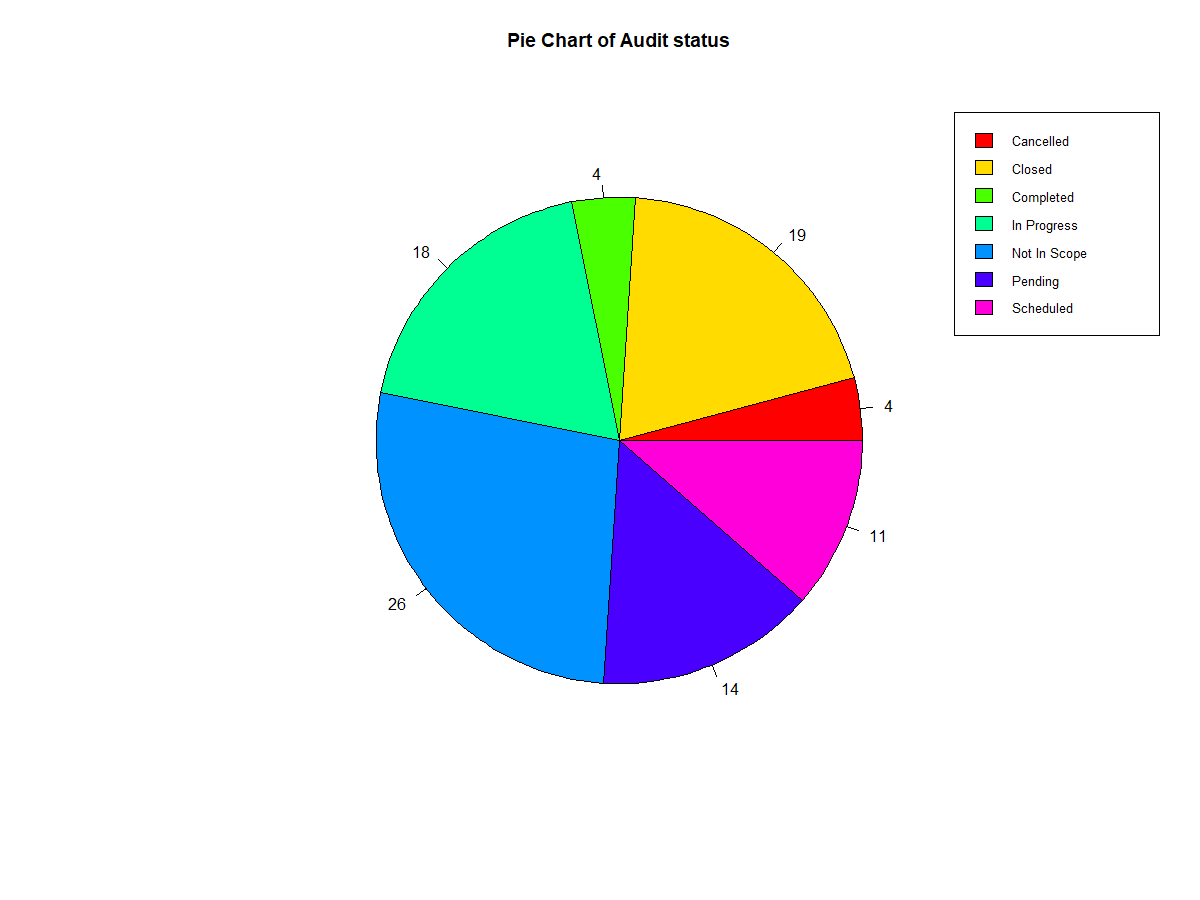
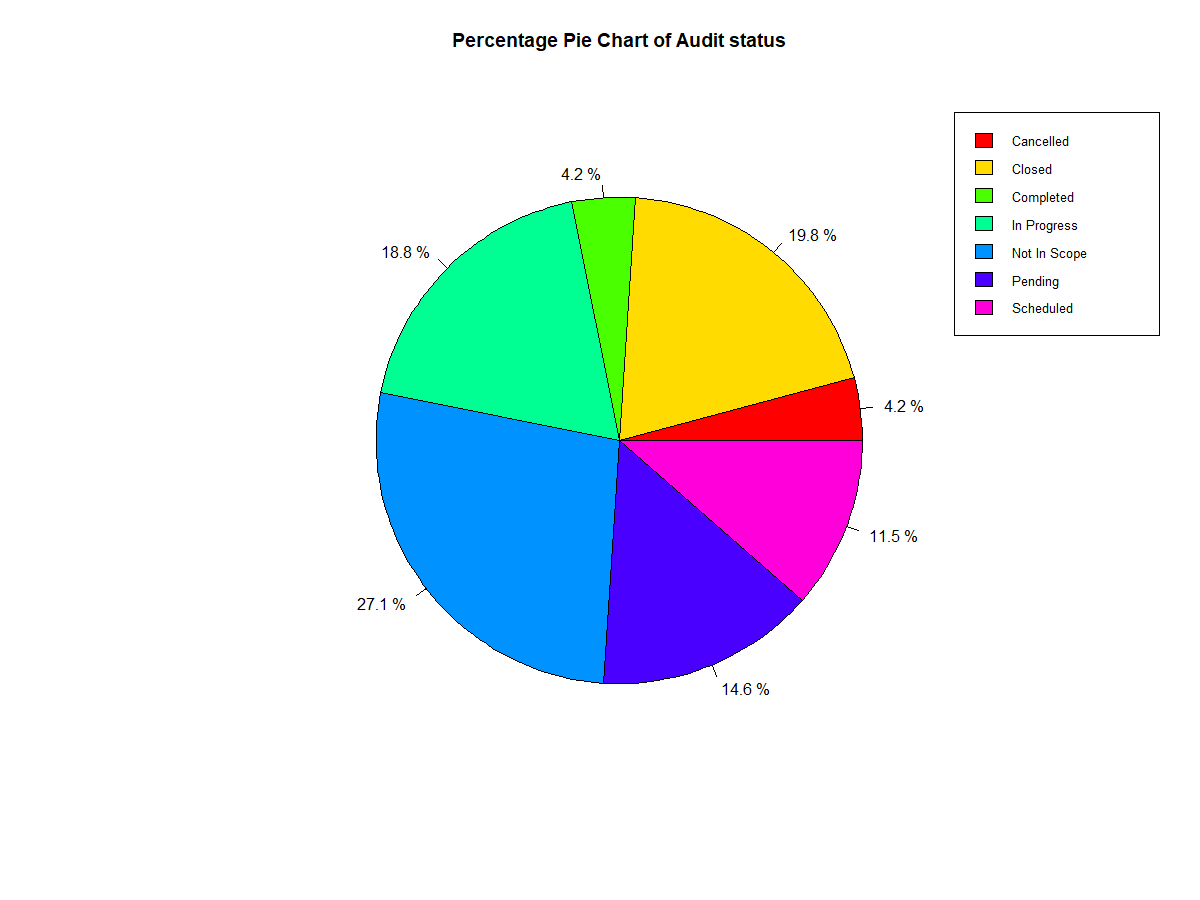
1. **Analysis**

**Part 1- GxP data**

The Excel File GxP includes information on 96 audits in 2017. Unfortunately, we cannot know whether data file is the full representation of the GxP audit in 2017 or just only a small fraction of it

* **Audit Status**





Approximately 1 out of 4 surveyed auditors are not in scope.

19.8 of audits are “closed” while the “completed” part accounts for 4.2%. People with no prior expereince might not know the difference between a “closed” and a “completed” audit. But since I have worked in the industry, I know that an audit is considered “completed” when they are issued by the lead auditor and considered “closed” when all planned audit activities have been executed or agreed by the audit client.

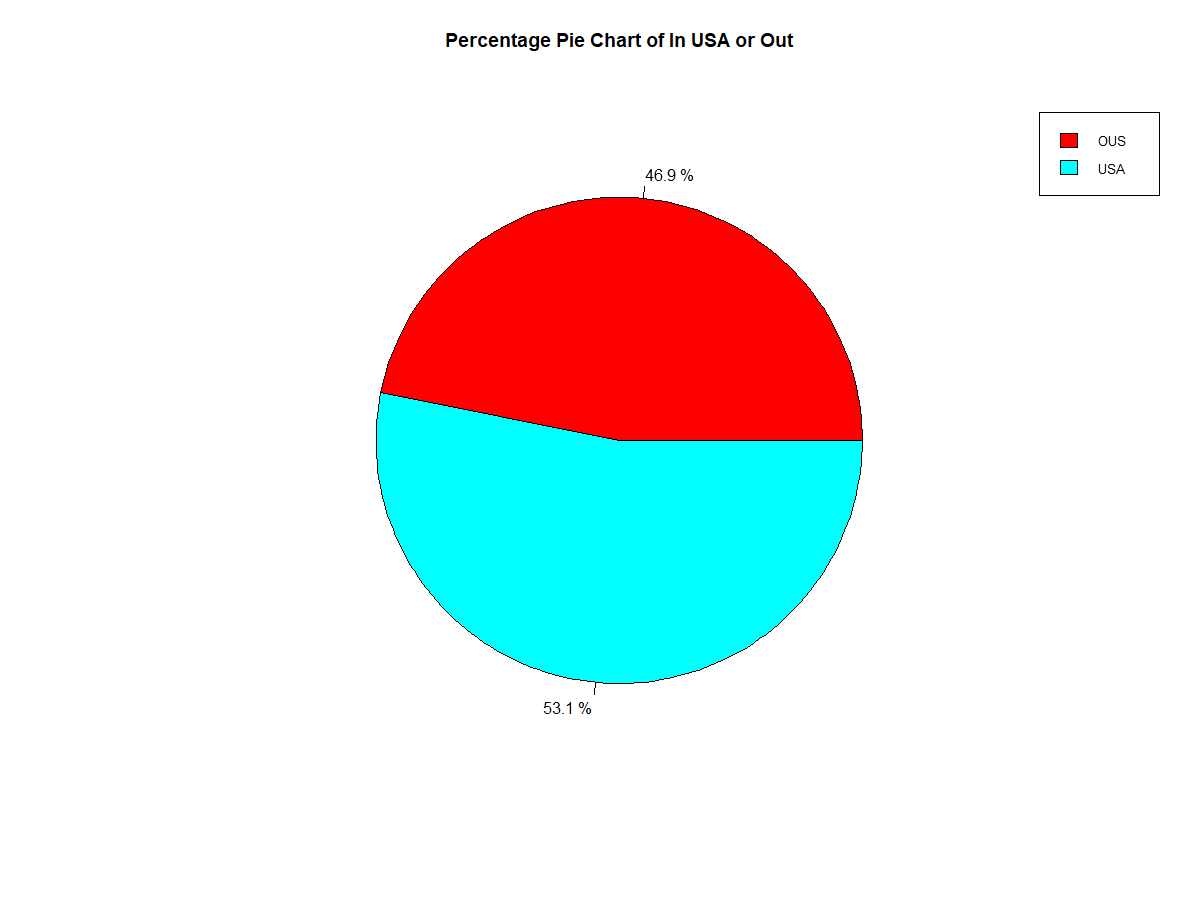
The ones are “in progress”, “scheduled” and “pending” account for 18.8%, 11.5%, 14.6%, respectively. It is noteworthy to know that pending audits do not have their dates of the audit process published.

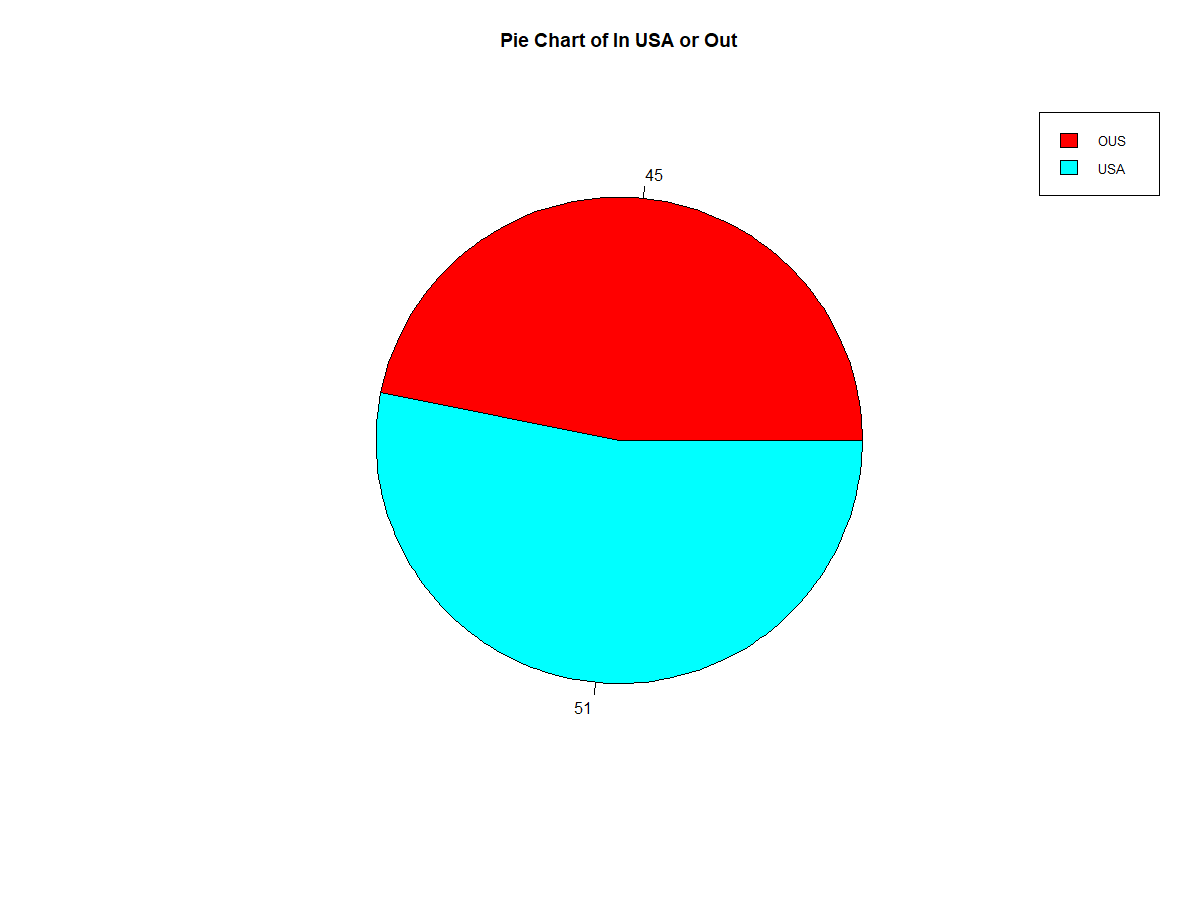
4.2% of the audits were canceled. There should be more investigation into the reason why the they are scrapped

* **In USA or OUS**

The numbers of auditors living inside and outside of America are approximately equal to each other. However, there are lightly more auditors living inside of America (53.1%) compared to the those living outside of America (46.9%).

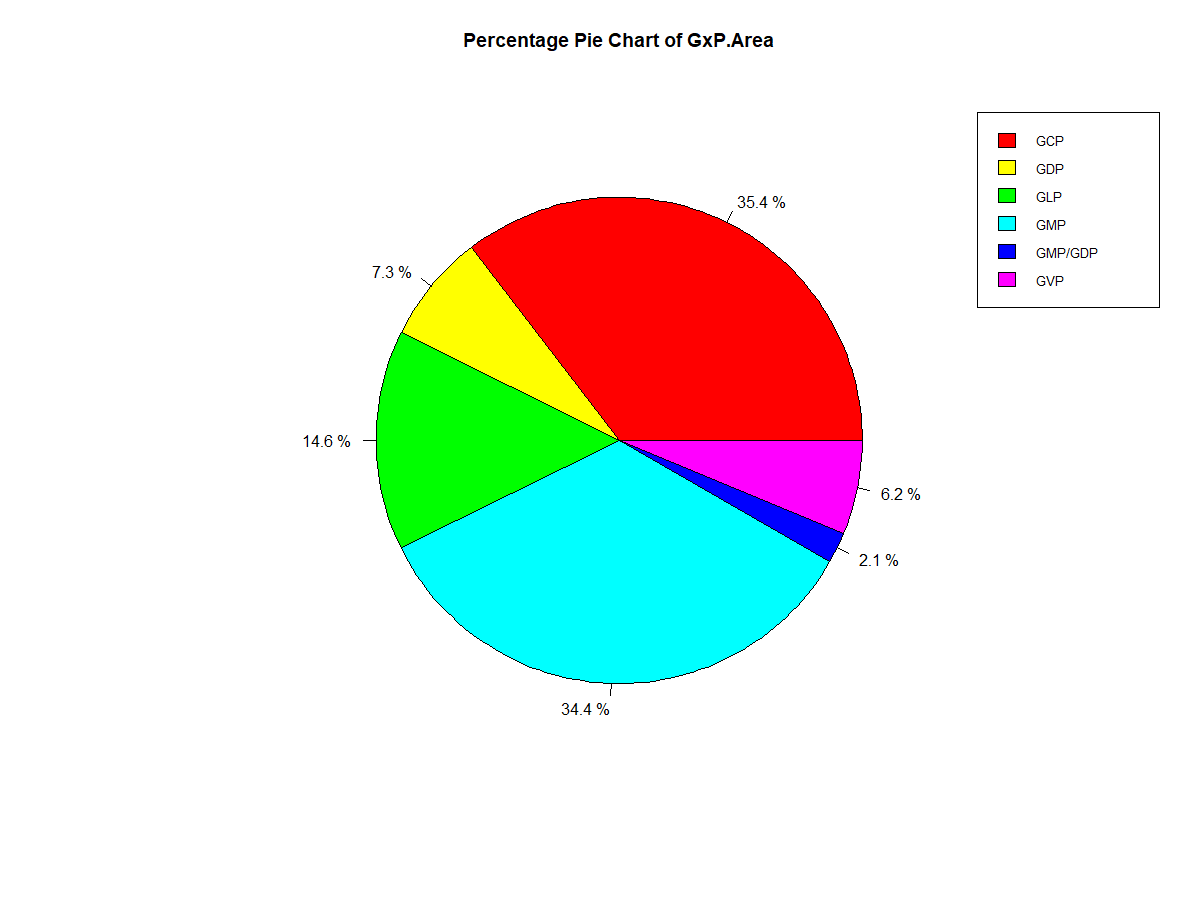
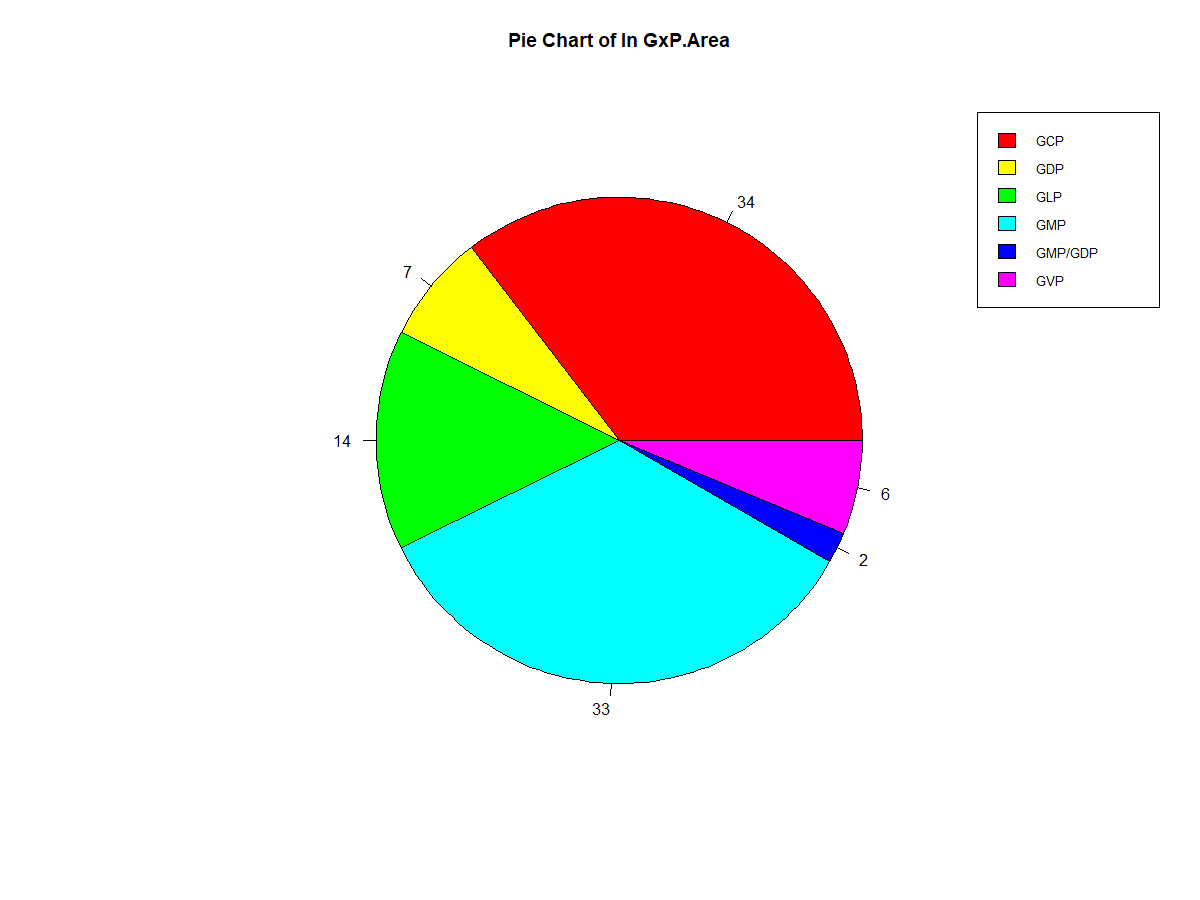




****

* **GxP Area UNCLEAN DATA)**



****

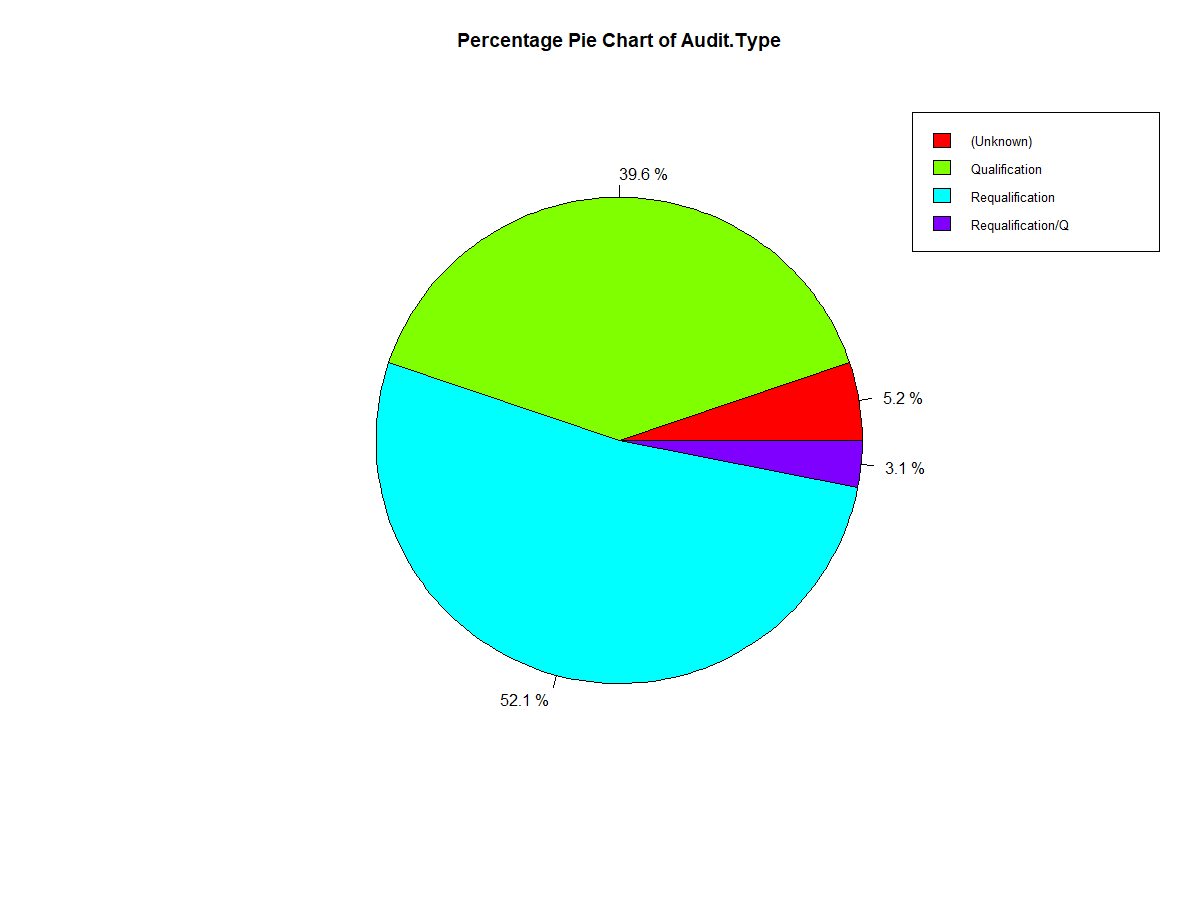
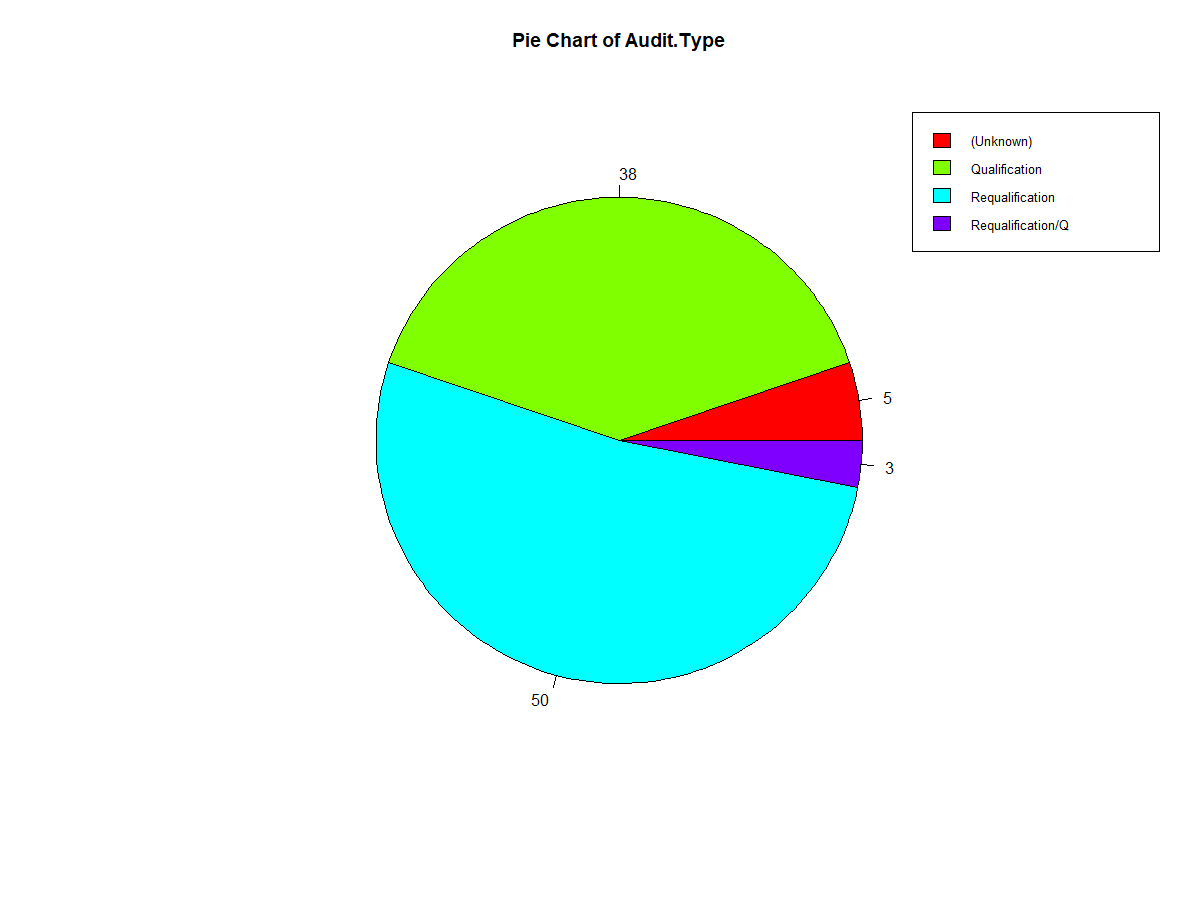
Approximately 3 out of 10 surveyed auditors have “Good Clinical Practice” .

34.4% of auditors are “Good Manufacturing Practice, audits of manufacturing sites.” while the “Good Distribution Practice, audits of distribution and logistic partners” part accounts for 14.5%. However, there is no clear indication of what is the difference between these two

The ones are “GMP/GDP”, “GVP” and “GDP” account for 2.1%, 6.3%, 7.3%, respectively.

* **Audit Type**



****

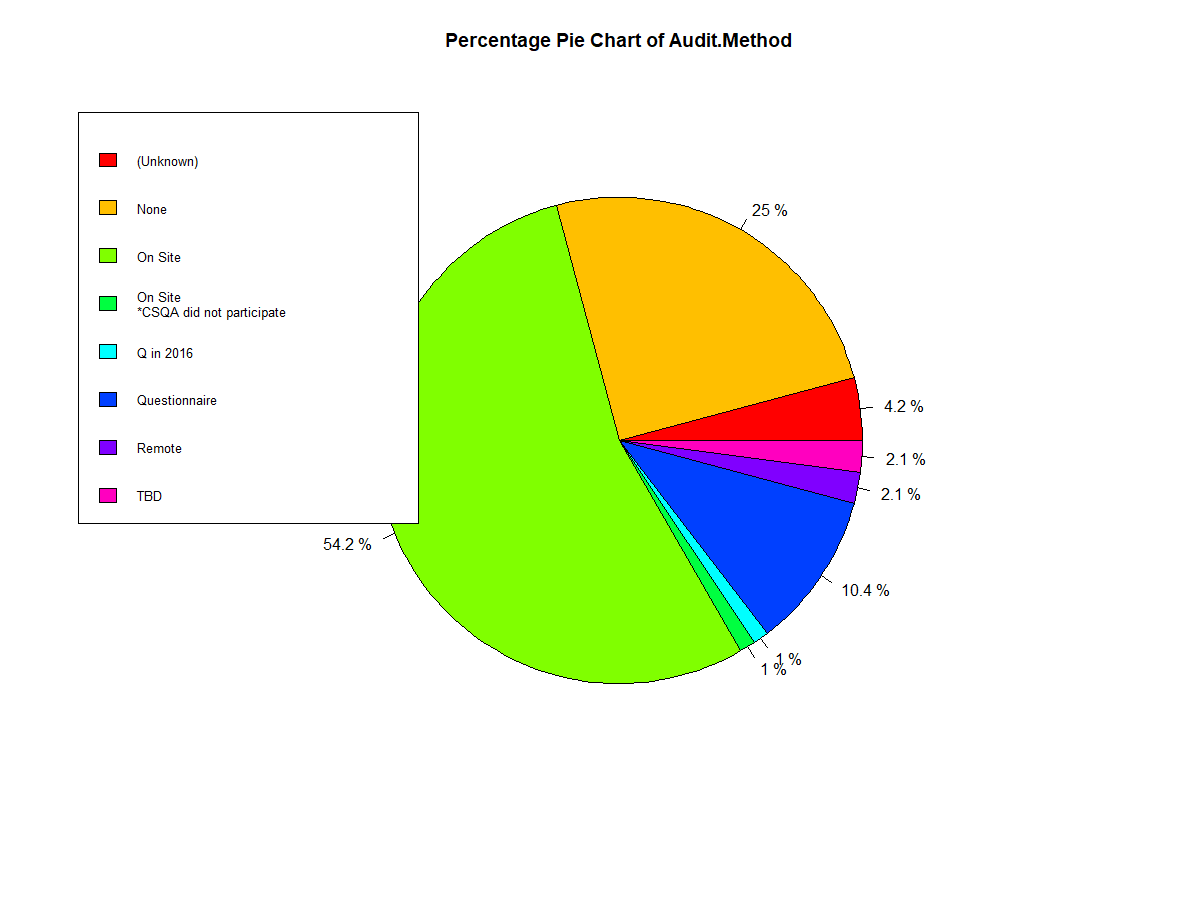
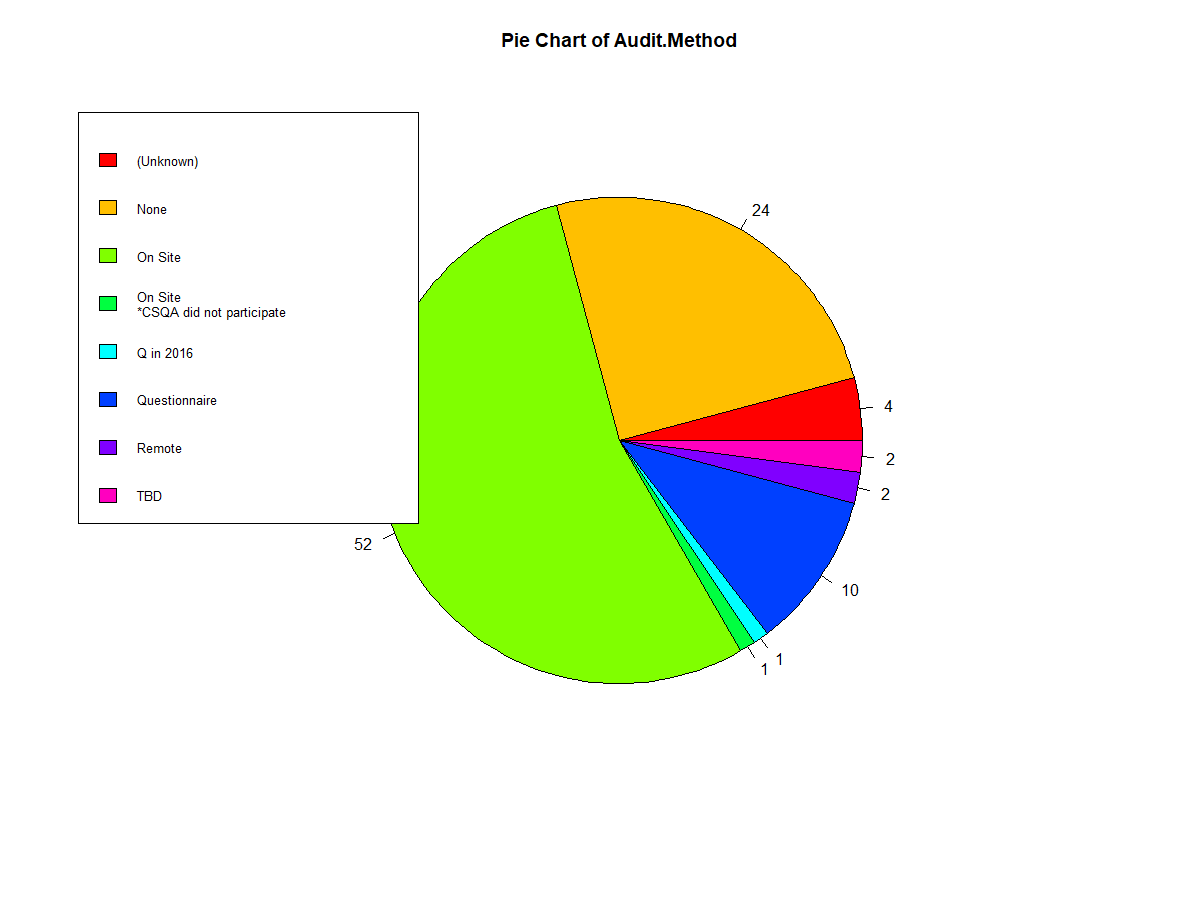
Approximately 1 out of 20 surveyed auditors ‘s types are not documented.

39.6% of auditors have “Qualifications” while the “Requalification” part accounts for 52.1%. We are not sure what the difference between “Qualification’ and “Requalification” is.

The ones with ”Requalification with Question “account for 3.1%,

* **Audit Method (UNCLEAN DATA)**



****

More than half of the surveyed auditors conducted auditing on site.

10.9% of auditors chose Questionnaire method while those who chose “None” account for 26.1%. I am not so sure what “none” means in this case.

The ones are “on-site \* CSQA did not participate”, “Questionnaire in 2016” and “remote” account for 1.1%, 1.1%, 2.2%, respectively.

2.2% of the audit methods were unknown or “TBD”.

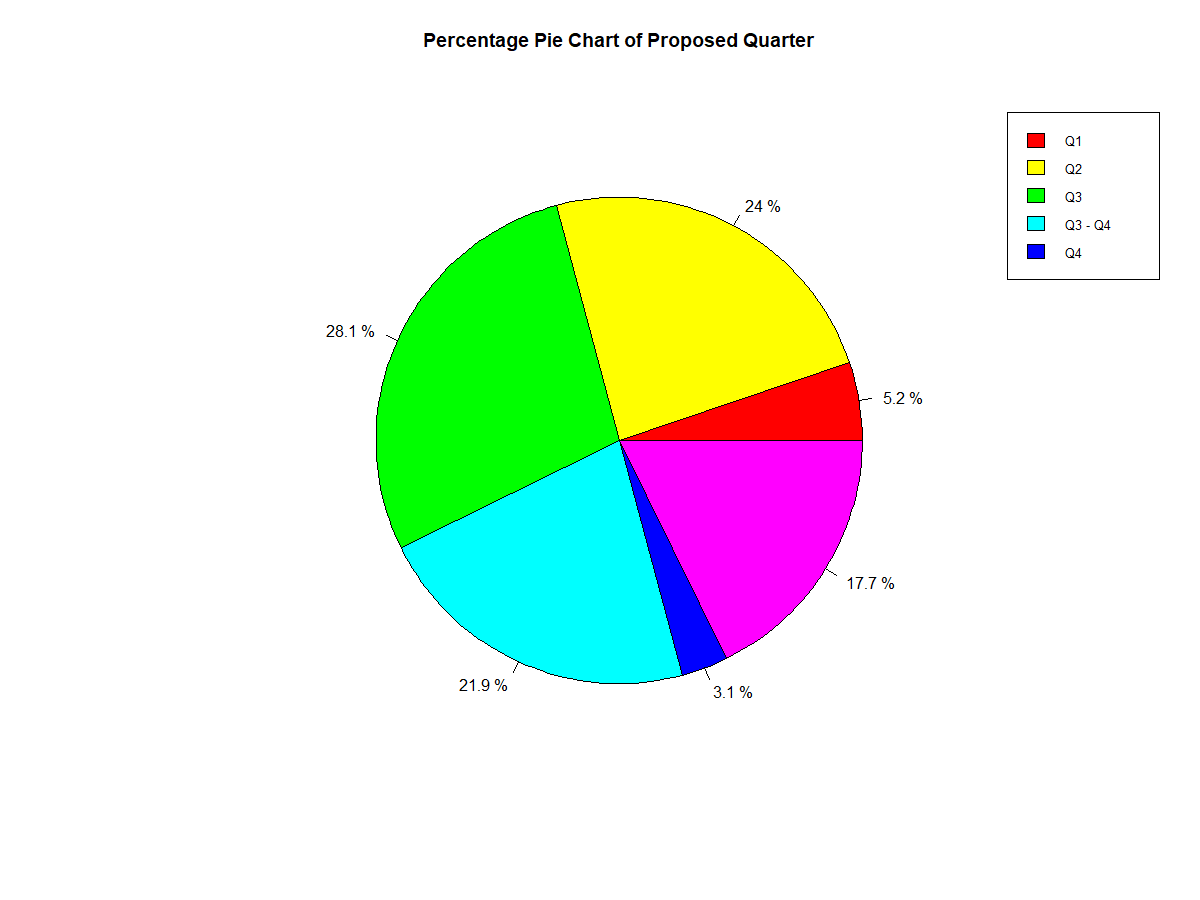
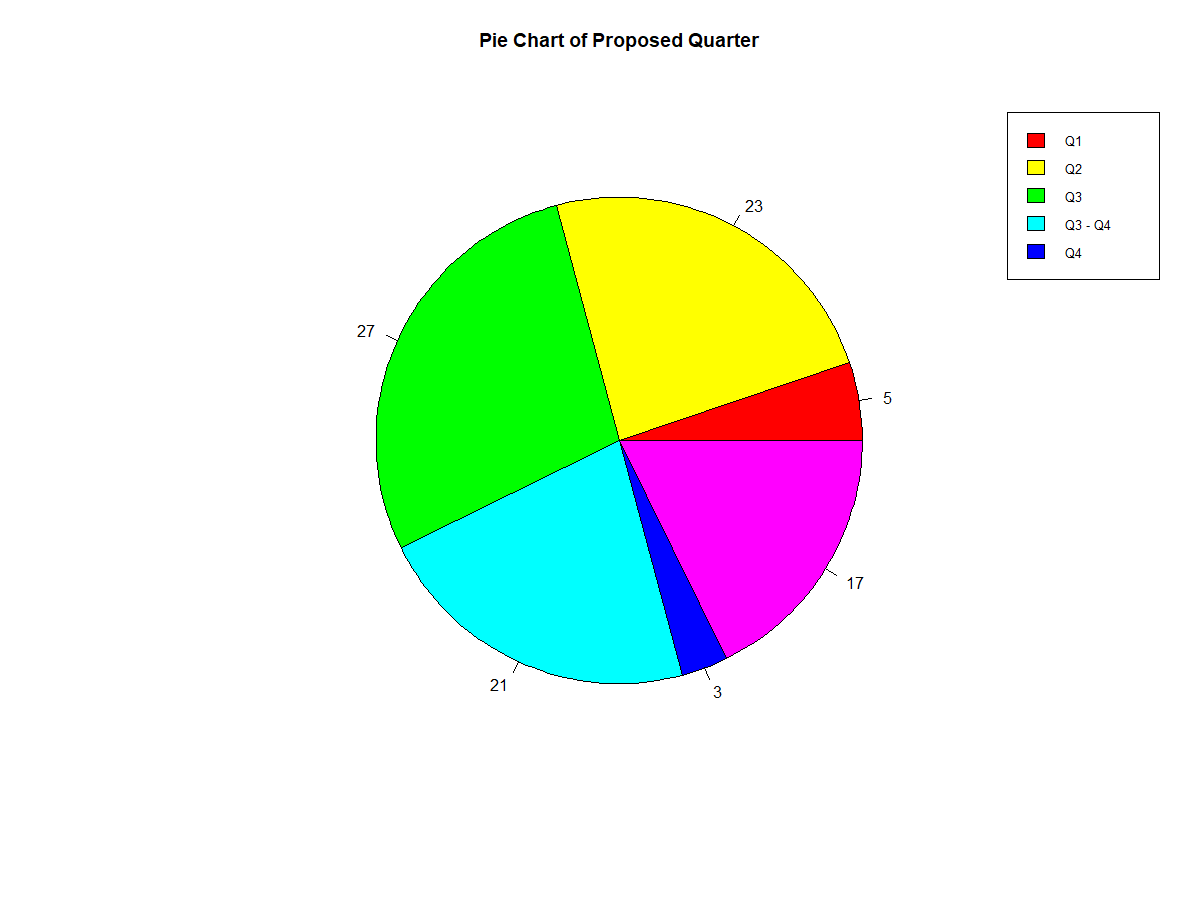
* **Proposed Quarter UNCLEAN DATA)**

Quarter 4 has the lowest percentage of total audits out of all the quarters with only 17%

The ones that belong to “Q1”, “Q2” and “Q3” account for 24%, 28.1%, 21.9%, respectively. It is quite equally distributed across the first three quarters.

Those cannot be classified take up 5.2% of the total and those are proposed to belong to Q3-Q4 have



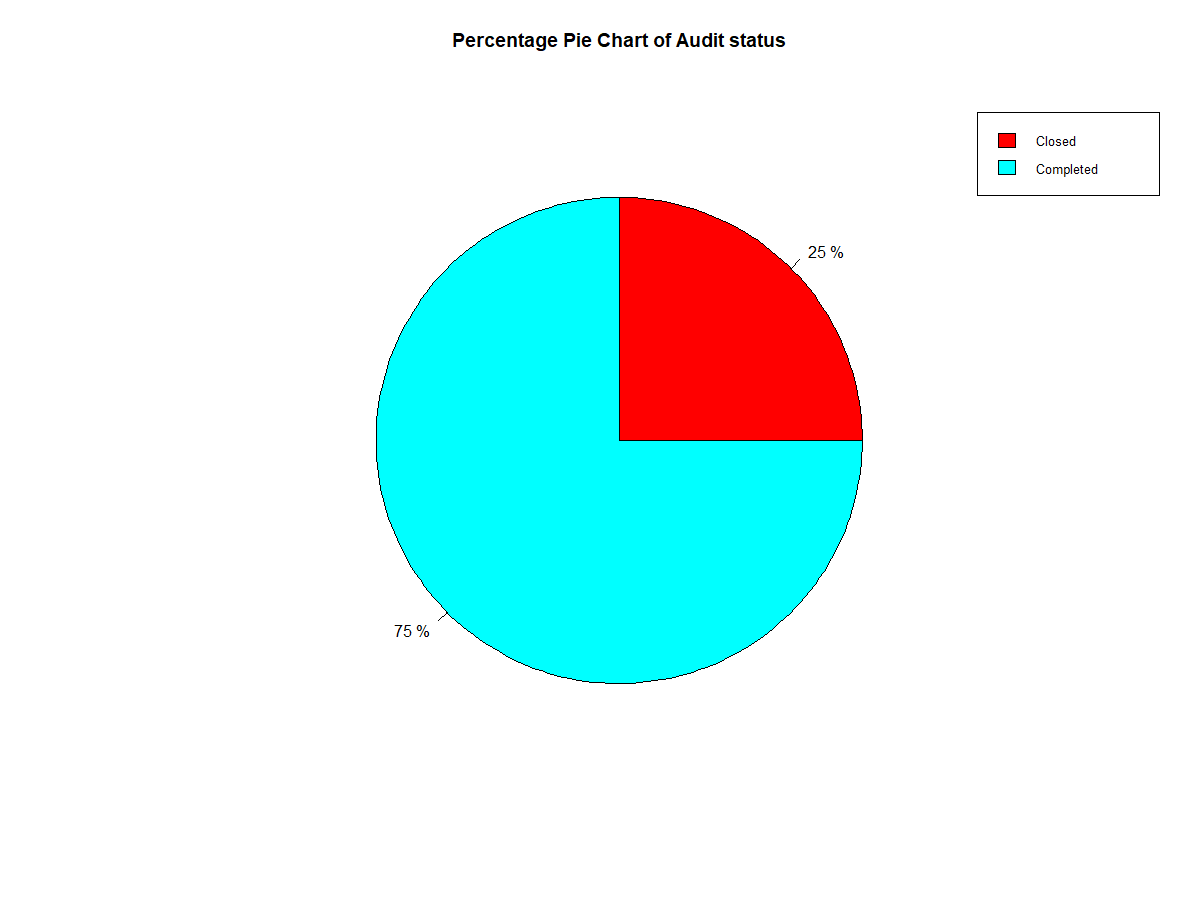
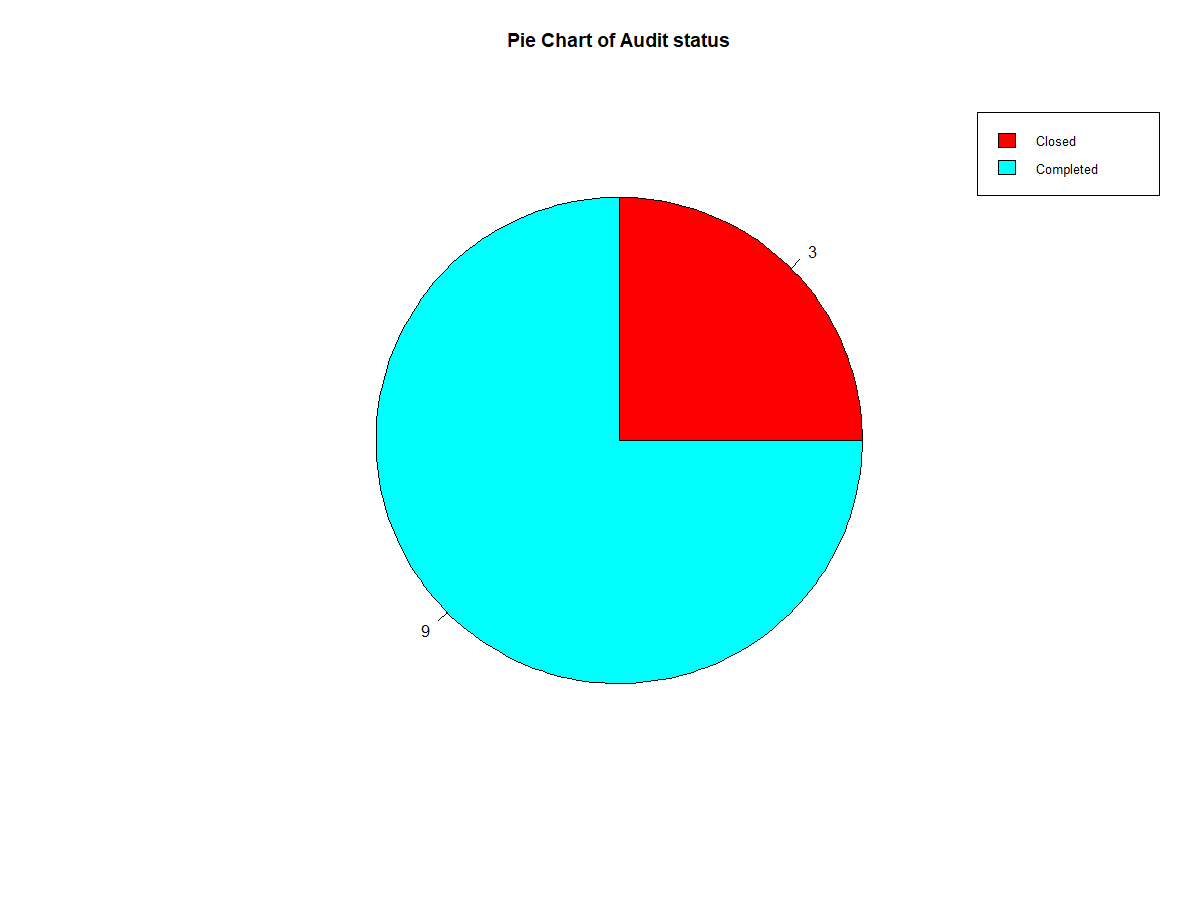
****

**Part 1- CSQA data.**

The Excel File CSQA includes information on 12 audits in 2017. It is basically a short version of the GxP file above, so there is much smaller groups of auditors to focus on.

* **Audit Status**

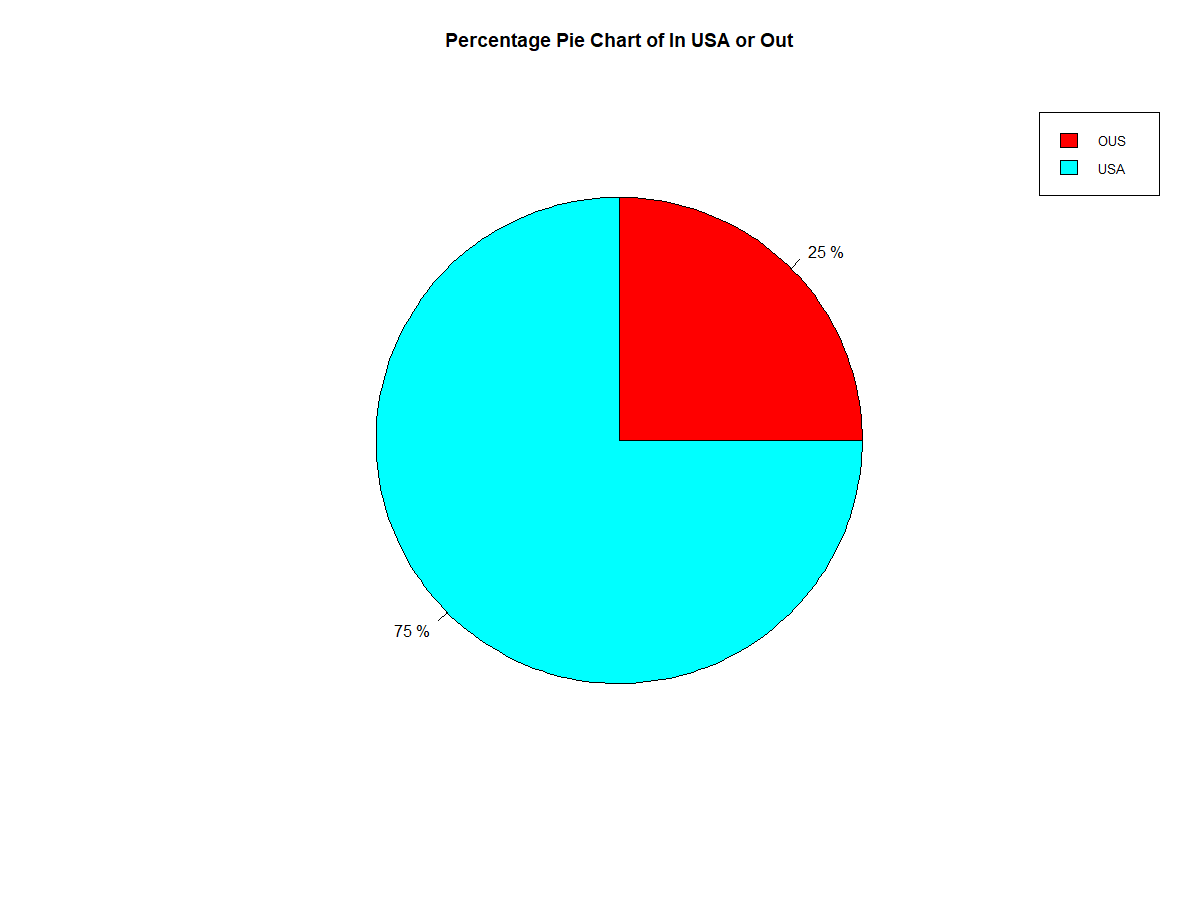
75% of this condensed groups audit’s works are “Completed” while 25% of them are totally “Closed” 

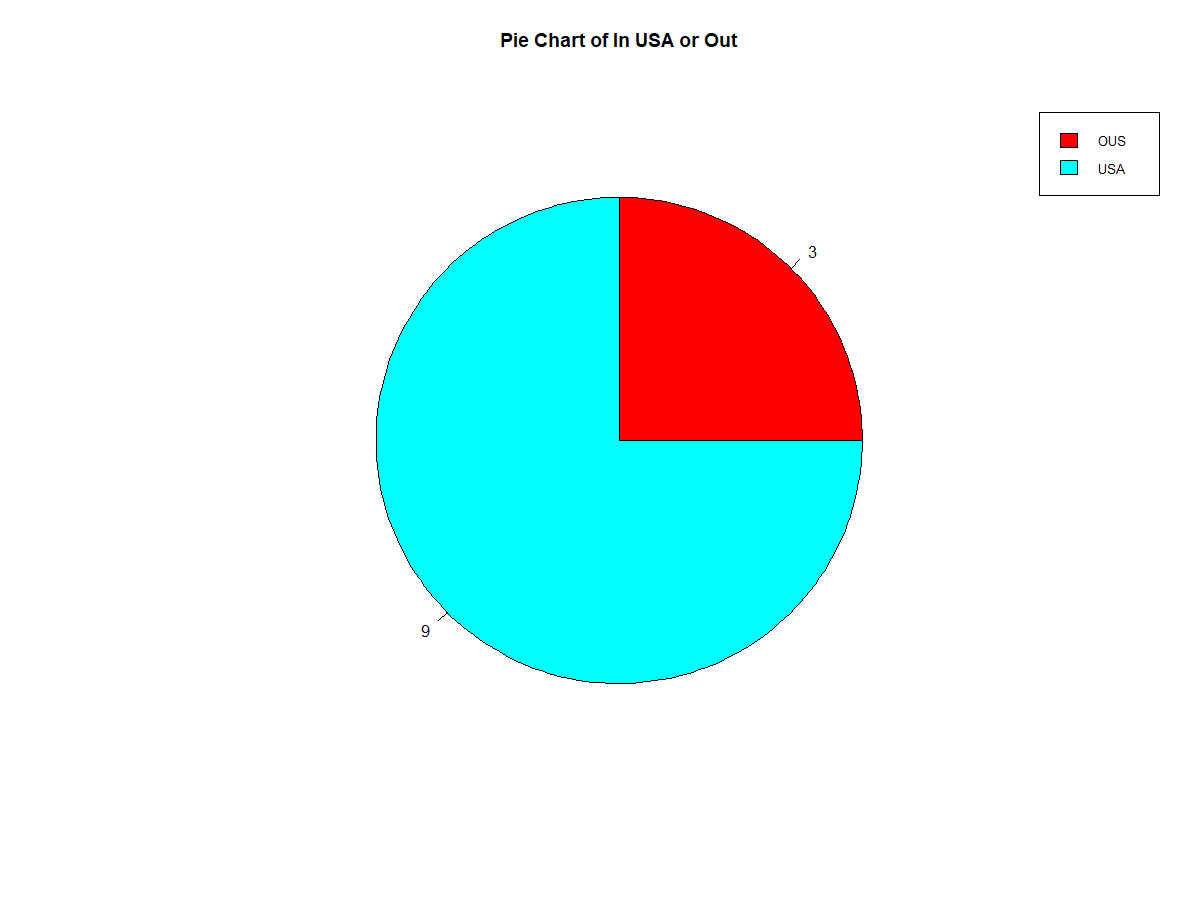


* **In USA or OUS**



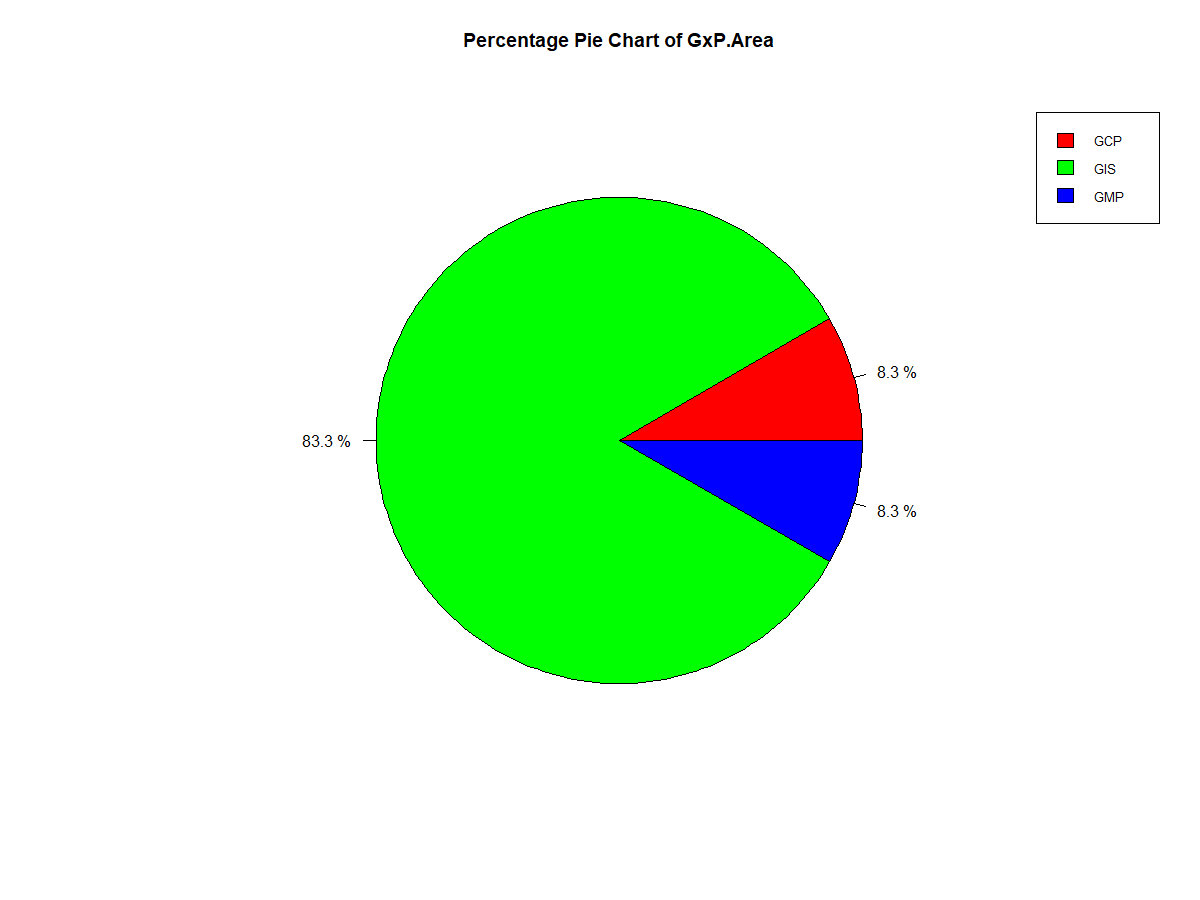
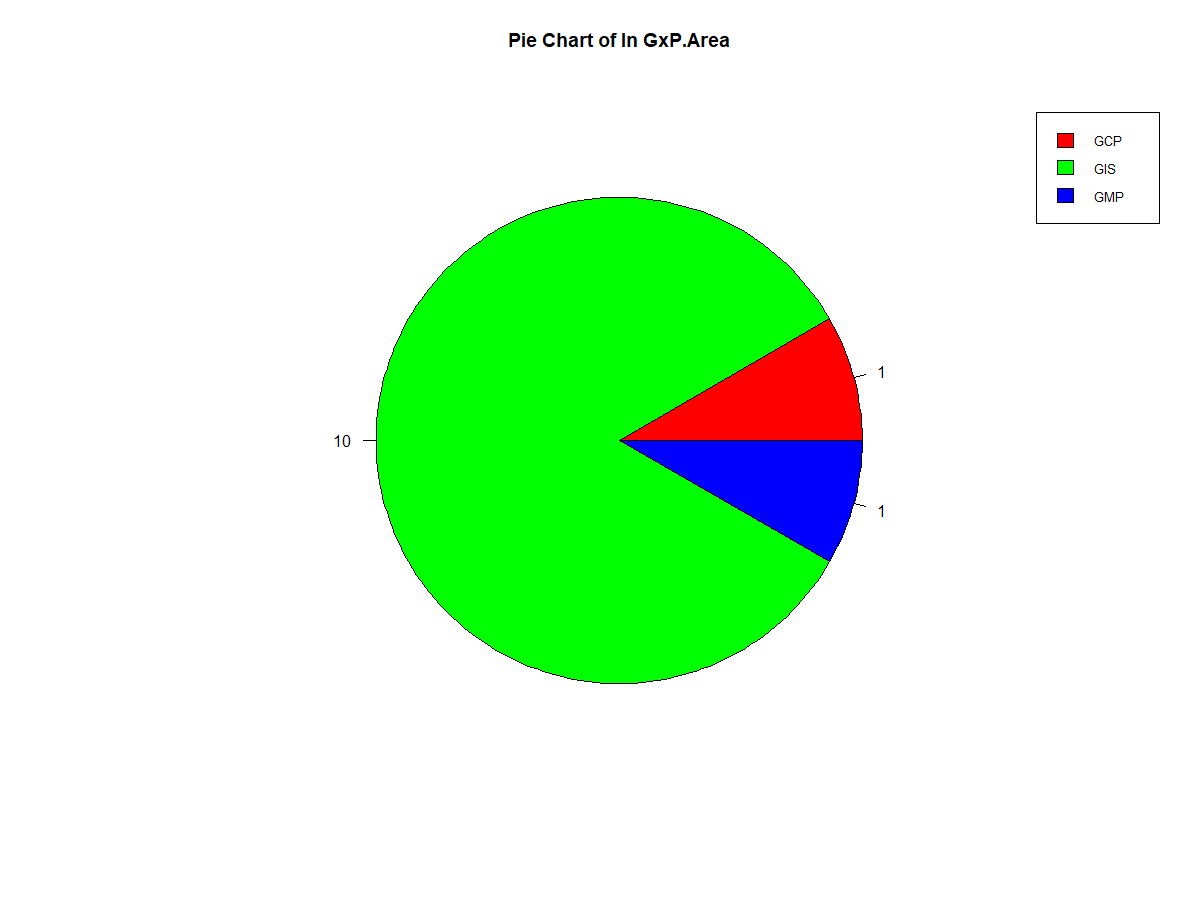
75% of this condensed groups audit’s works are from inside of the USA while 25% of them are from outside of USA



****

* **GxP Area**

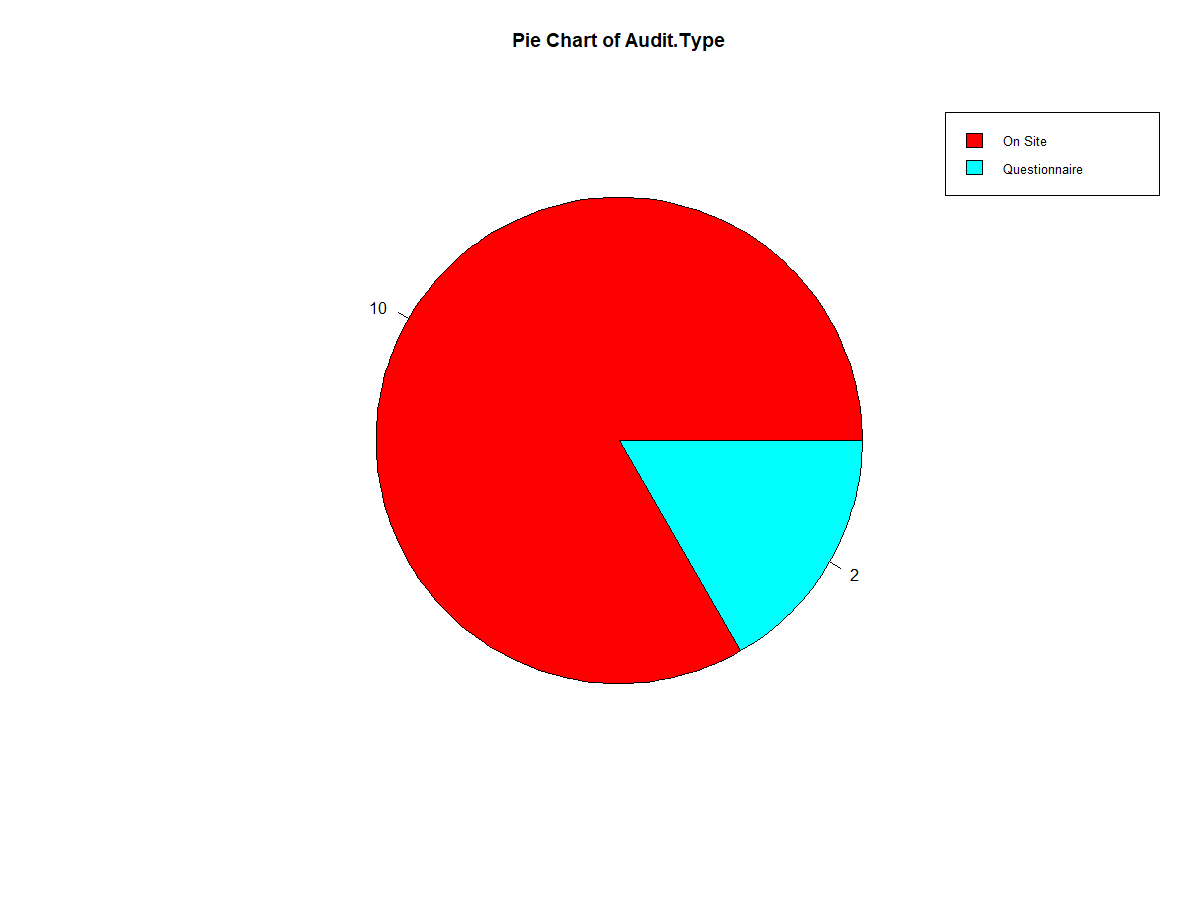
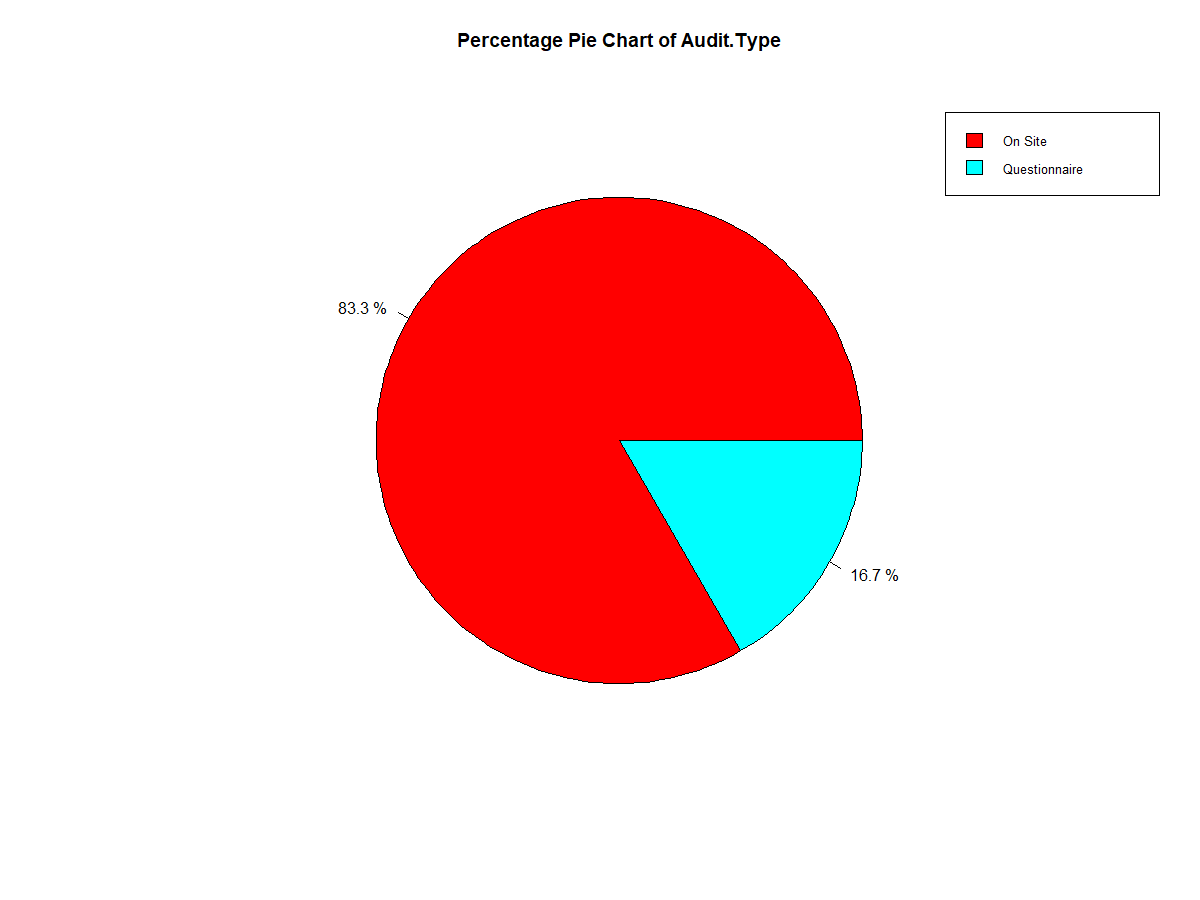
83.3% of this condensed groups audit’s works belong to the “GIS” Area while those in “GMP” and “GCP” both account for 8.3% each.

****

* **Audit Type**

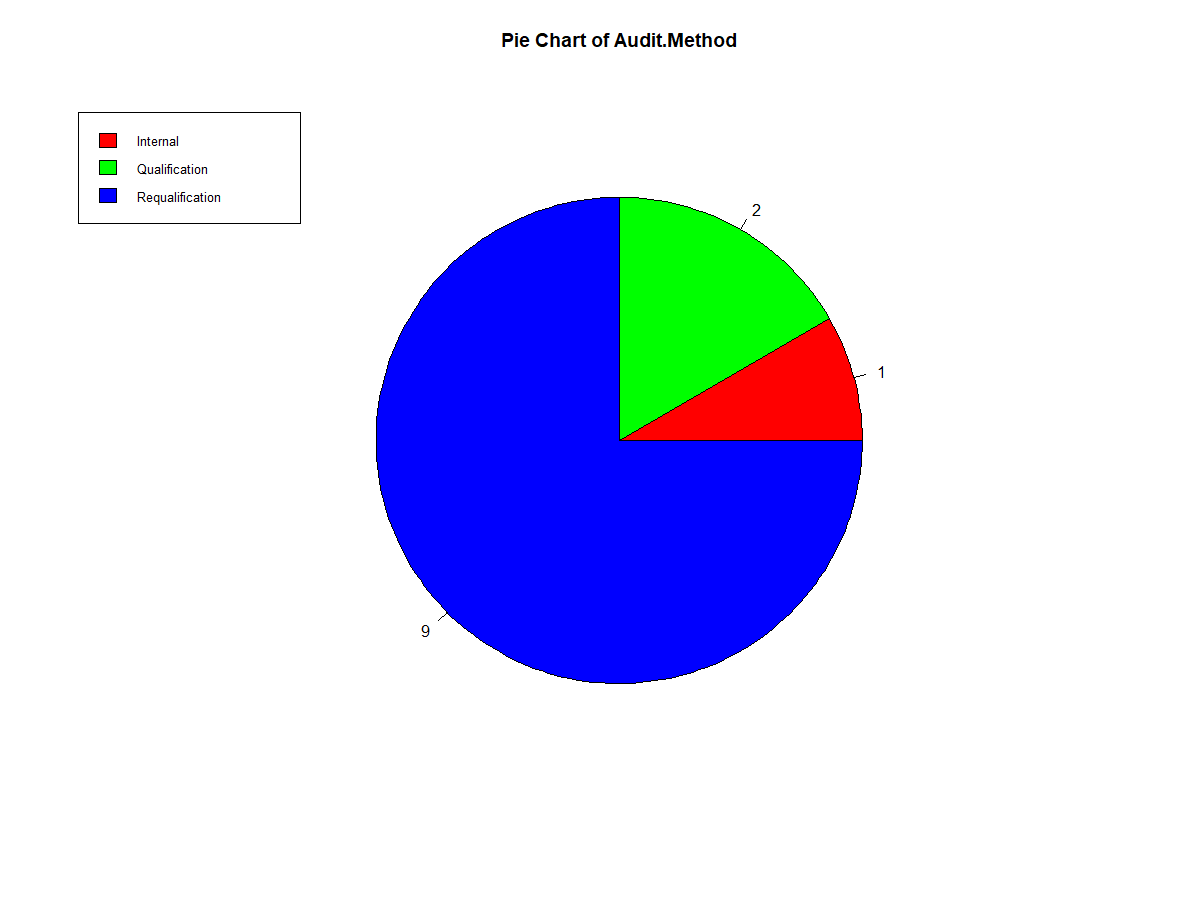
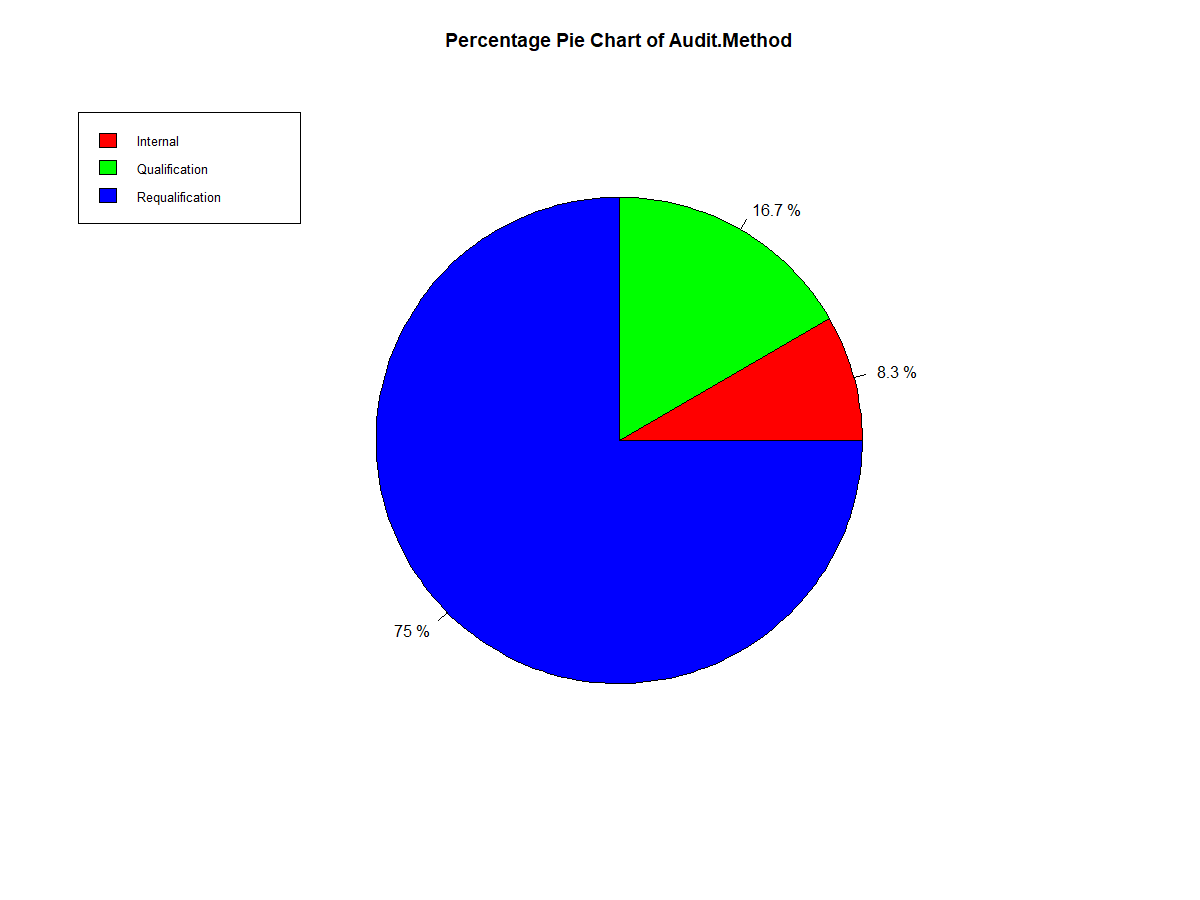


83.3% of this condensed groups audit’s works are conducted “On Site” while 16.7% of them are conducted through “Questionnaire”

****

* **Audit Method**

75% of this condensed groups audit’s works are conducted through “Requalification” while 16.7% of them are executed through “Qualification ” and only 8.3% of them are done through infernal method

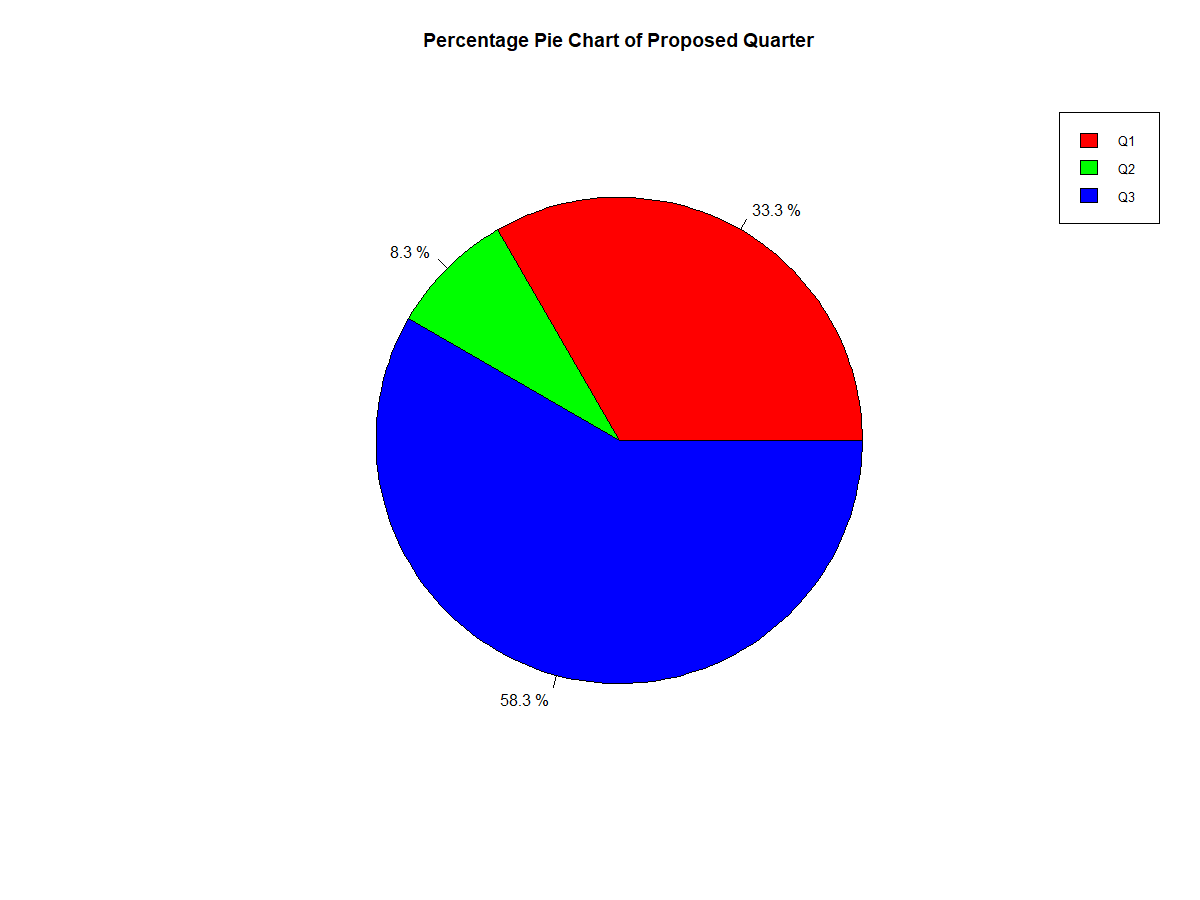


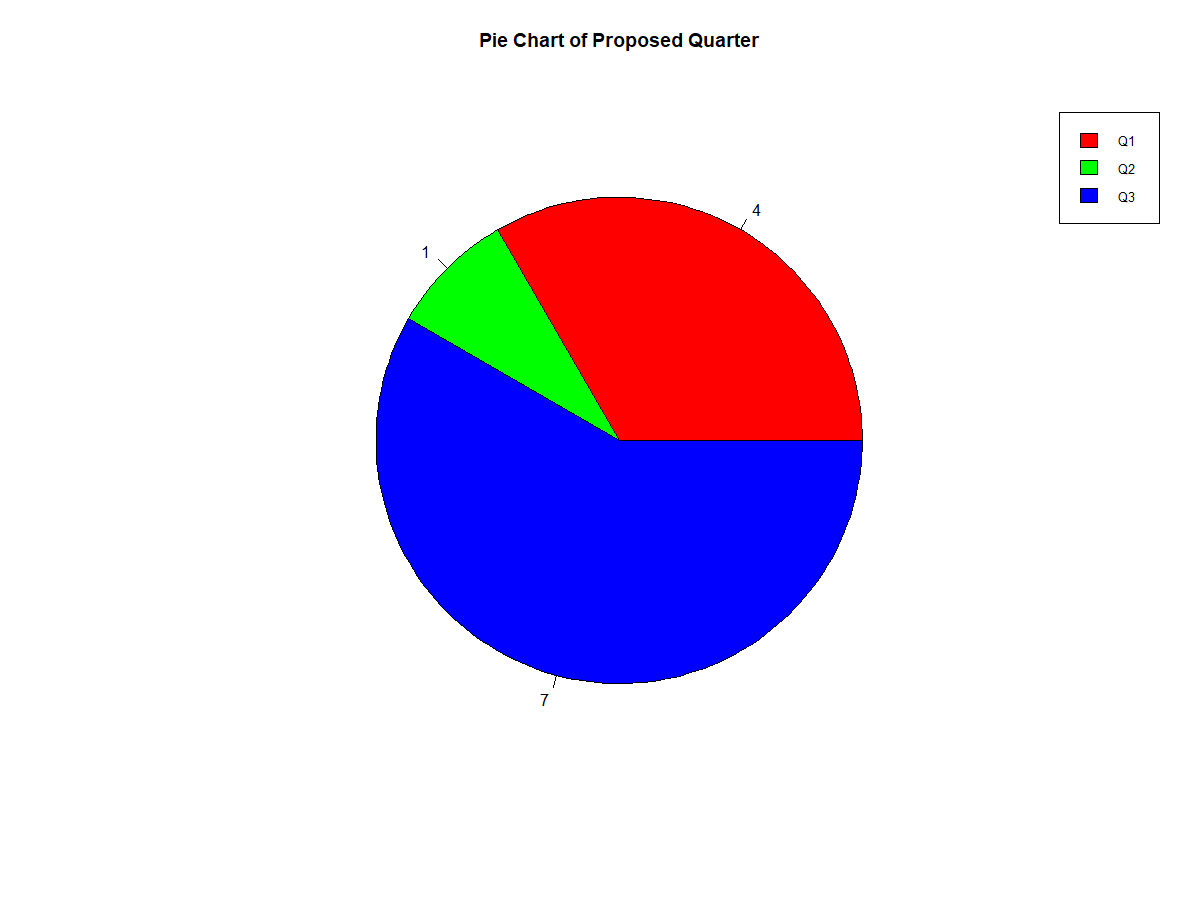
”Requalification

* **Proposed Quarter**



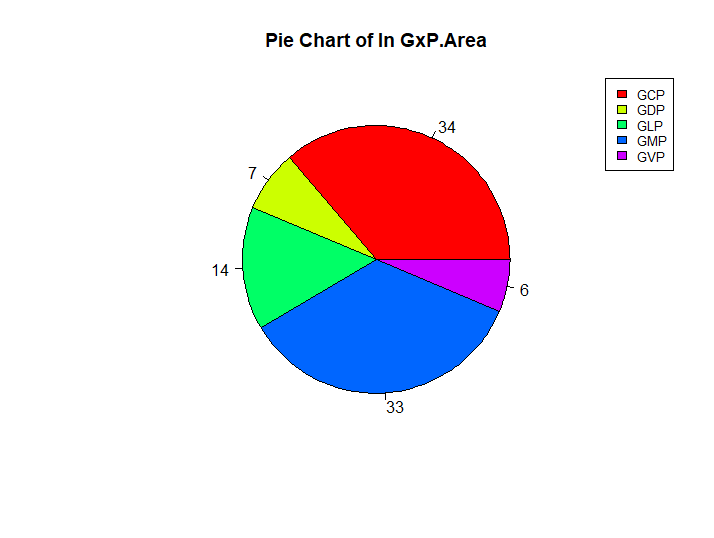
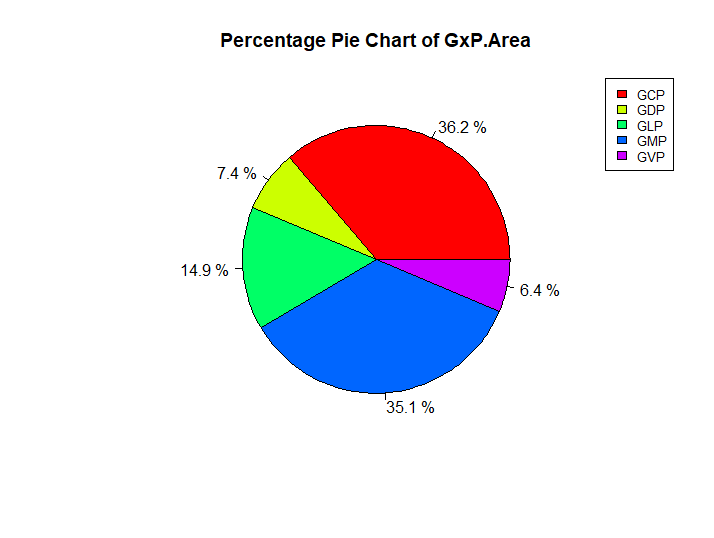
58.3% of this condensed groups audit’s works are belong to “Q3” while 33.3% of them belong to “Q1” and only 8.3% of them belong to “Q2”





It is noteworthy to know that there are some out-of-range variables in the GxP, Proposed Quarter and Audit Method variables of the GXP audit file that needs to be addressed because they variables exceed the Data Type’s limitation. For example: GxP only uses CHAR (3) data but there are inputs of “GMP/GDP” that the machine cannot recognize. So here are the results again:

* **GxP (CLEAN)**

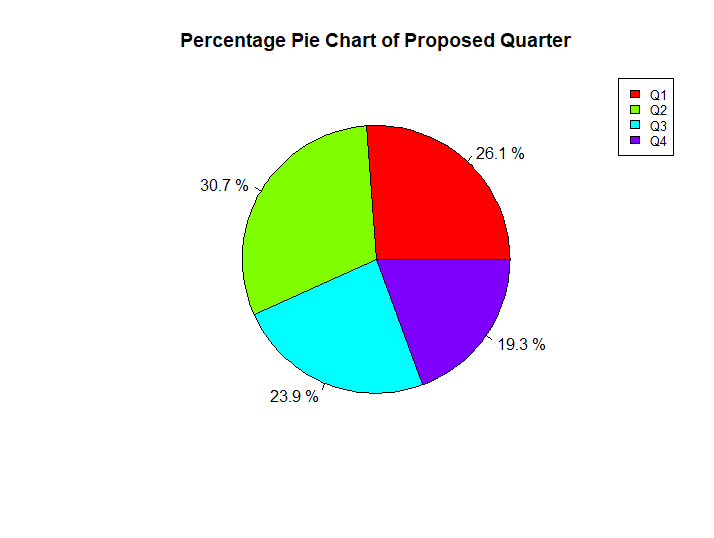
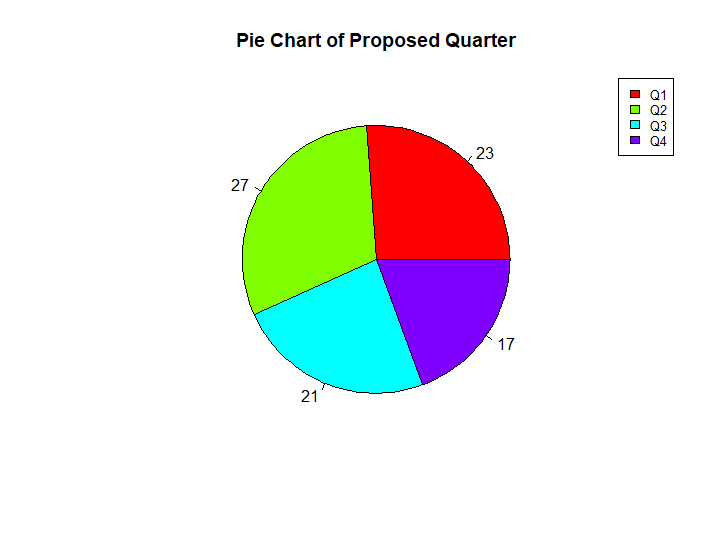
****

The

The results are much easier to interpret compared to the unclean data version. We see that GMP and GCP are the dominant groups that account to 70% of the total population while the others account for less than 30%

* **Proposed Quarter (CLEAN)**

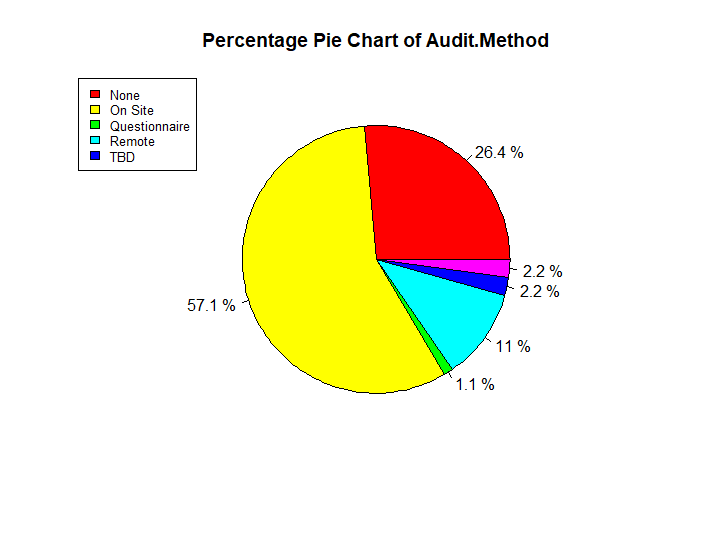
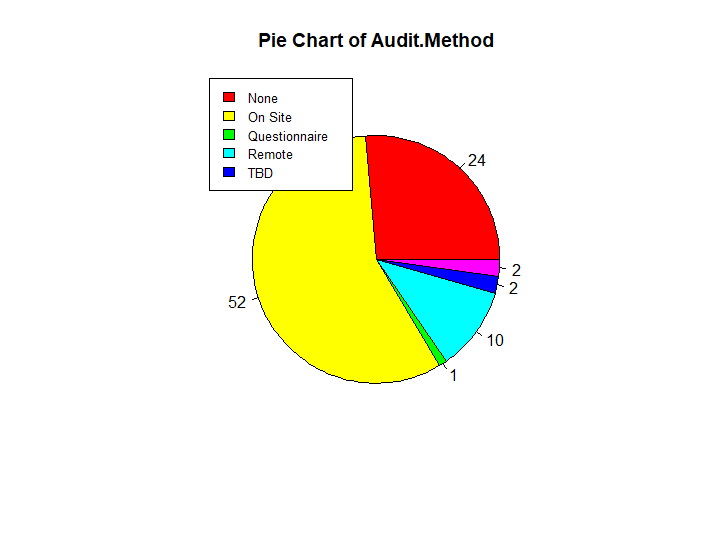


****

Compared to the unclean data, all of the quarters have approximately the same percentage with Q2 being the most popular (30%) and Q4 being the least likely quarter to be proposed.

* **Audit Method (CLEAN)**



****

Since we only take out one out-of-range variable. The result would not change that much

**Part 2**



For Days\_Intake\_QSent, while GXP audit has the mean and median drastically different from each other (17.4 and 2),CSQA audit does not (50.5 and 50.5).



Days\_QSent\_QReceived represents the differences between the day the questionnaire is sent and the day it is received. In general, a questionnaire will take around 30 days to reach the receiver. However, according to the computer, 18 is the number of days that it should take for the questionnaire to arrive. It is unknown that the rule applies to all the questionnaires or not because as we can see that we can only get the information of 10 variables out of 96 from the GXP Audit file as well as the aforementioned wide dissimilarities between mean and Median of GXP with CSQA’s ones . There might be technical complication that prevents the dates’ documentation



Days\_OnSiteScheduled\_AuditStartDate illustrates the number of days difference between the supposed on site schedule and the start day of auditing. The number is quite great and somewhat similar to each other on both reports. This means that they are very late in carrying out the task according to the schedule, almost 2 months. Since the variable is around 50% and the mean and the median of GXP almost equals to CSQA audits, I would say that it will likely apply to all.



For Days\_StartDate\_EndDate , the mean and median are approximately similar to each other regardless of the different in number of appropriate variables (50 and 9).



Days\_AuditEnd\_FinalReportDue” , the mean and median are approximately similar to each other regardless of the different in number of appropriate variables (50 and 9).



Days\_FinalReportDue\_CompletionDate , they have the same median at 0 across two excel documents but CSQA has a higher mean compared to the GXP’s one (7 and -0.2, respectively)

, the mean and median are approximately similar to each other regardless of the different in number of appropriate variables (50 and 9).

**Part 3**

From a respective point of view, I would not merge the two files together because there are not enough evidences that they would go well together. GXP reports realistic results from 96 variables that are recorded while CSQA is the computer’s selected ones from a sample of 12 only. Furthermore, there are only 3 out of 6 variables from part 2 that yield the equal mean and median, which means that if we put them together, there will be unpredictable changes in the outcomes. Therefore, it would be safer to keep them all separate.

The company is doing not a good job in keeping track with the scheduled dates because the projects did not start at the right time which led to the overdue projects way beyond the deadlines. They need to improve their planning before carrying out the task and take other elements into mind like risk, equipment,etc. (“What makes a good production plan?,” 2019)

1. **Conclusion**

To sum up, we have seen the many comparisons between two files and learnt that there is a wide gap between the GXp and CSA files. It would be better to have a complete set of data wit all the parts being completed according to the Data Directory so that we can do the analysis more precious and merge them together. (Mathara, Choy, & Madhushani, 2015).

**References**

Mathara, S., Choy, S., & Madhushani, A. G. I. (2015). Quality assurance and quality control in ERP systems implementation. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, *11*(1), 70–83.

What makes a good production plan? (2019). Retrieved from https://www.bdc.ca/en/articles-tools/operations/operational-efficiency/pages/production-plan-top-tips-improving-operations.aspx