12/5/2022

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

IE 266 ENGINEERING STATISTICS

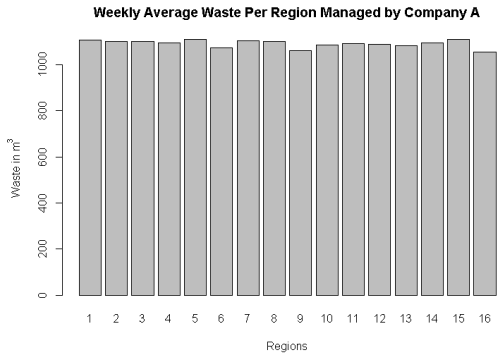
Case Study 1 Report

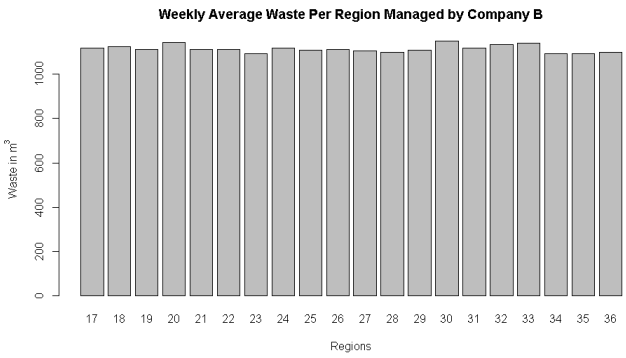
GROUP 25

Abderrahmane Harkat	2415297
Youssef Nsouli	2487494

Q1-a) Use appropriate descriptive statistics (graphical and numerical tools) to comp…

Comparing the weekly workload of regions for each company, the following bar graphs are produced:





Figures 1 and 2: Bar graphs of weekly average waste for each region (in m^3) managed by each of companies A and B.

Analyzing both plots, the data and distribution for all regions is fairly uniform, and can be approximated as such.

When comparing the weekly workload for each company, it is observed that company B puts more effort into collecting waste, since they collect, on average, about 18430 m^3 of waste weekly, as opposed to company A's 14377 m^3. However, company B manages more regions, so it is fair to analyze the waste collected for each region weekly.

Since the first company has an average of 899 metric tons of collecting recyclable waste while the second one has an average of 921m^3, there's no reason to doubt that any company spends more effort collecting recyclable waste.

In addition, when analyzing the separation of non-recyclable waste and comparing the recycling ratio, there is still no reason to doubt that any company puts more effort than the other, since the ratios of recycled total waste to total waste are close: 0.824 for company A, and 0.826 for company B.

Q1-b) Is it possible to say that weekly plastic, glass, and aluminum wastes are correlat…

Calculating the correlation with R using the following function:

cor(|*waste type 1*|, |*waste type 2*|, method = "pearson", use = "complete.obs")

a table of the correlation of waste types collected by each company is produced:

Correlation Of:	Company A	Company B
Plastic and Glass	-0.16867641	0.25812971
Plastic and Aluminum	0.15883011	0.06749558
Glass and Aluminum	-0.04352241	0.24339488

Table 1: Correlation of waste types in the regions managed by companies A and B.

Yes, it is possible to say that some of the wastes are correlated, plastic and glass are negatively correlated for regions under company A as in the more plastic is consumed, it is observed that less glass is consumed, while for the regions under company B, plastic and glass are positively correlated which means the more plastic is used, the more glass is used as well.

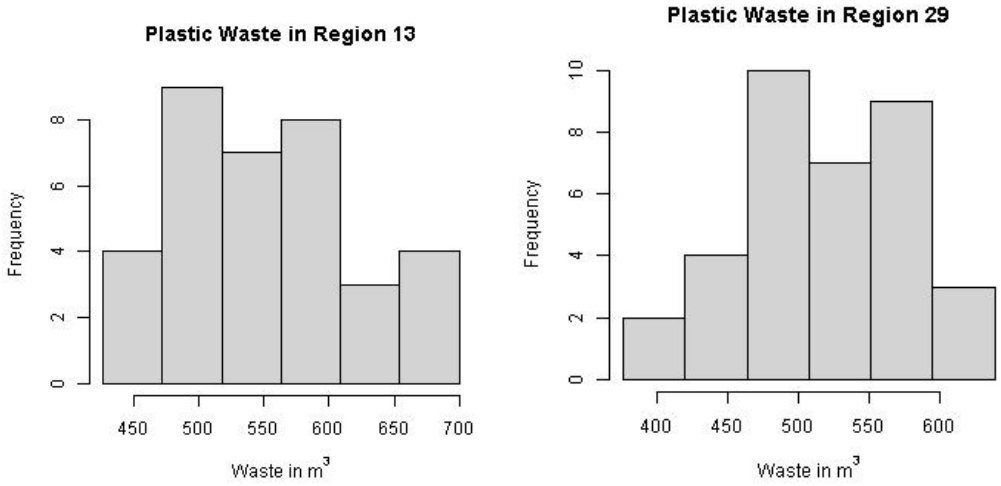
Plastic and aluminum are positively correlated for regions under company A's management as are glass and aluminum wastes for regions under company B, but for the rest of regions the correlation (plastic and aluminum for regions under B, and glass and aluminum in regions under A) is too low (close to 0) to draw a conclusion. Hence, it is safe to say there is no correlation between them.

As an additional remark, any waste type in any region that shows any sort of correlation shows *weak* correlation. All the absolute values of correlation are under 0.3 (Table 1).

Q1-c) Draw appropriate plots to visually check the normality of weekly plastic waste i…

The histogram for plastic waste and total waste for each region is plotted. In addition, the histogram for the total waste collected by each company across all regions is plotted as well. The total is 74 plots ($2 \times$*no of regions* $+$*no of companies*), and all plots can be found in the folder $Q1cPlots$.

Analyzing the plots it can be seen that the majority of the plots of weekly plastic waste are not normal (like Regions 13 and 29; figures 3 and 4), with some exceptions (Region 2; figure 5).



Figures 3 and 4: Examples of histograms that are not normal.

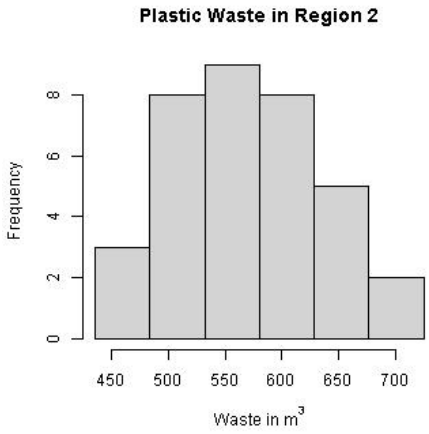
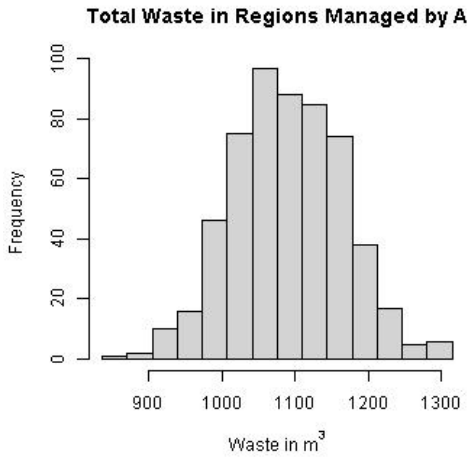
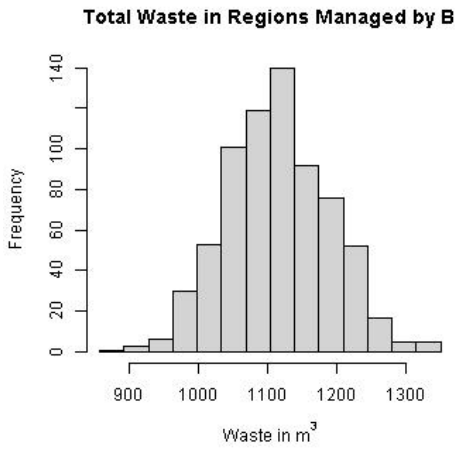


Figure 5: An example of a histogram that can be approximated as normal.

The same pattern is followed by the total waste in each region; some are normal, most are not, although there are more approximately-normal plots than plastic ones, perhaps due to the larger data size. Examples can be seen in Appendices A1 to A3. So, generally, normality of data cannot be assumed.

As for the plot of the total weekly waste across all regions, the plot for the regions managed by company A cannot be approximated as normal, on the other hand, the weekly total waste of regions managed by company B is approximatively normal.





Figures 6 and 7: Histogram of the total weekly waste across all the regions managed by each company.

Q2-a) Compare the average weekly amount of plastic waste collected by Company A wit…

Plastic Waste Collected by Companies A and B.

To compare the plastic waste collected by both companies A and B, it is imperative to first determine whether the variances are identical for both companies. To find the variance, it is necessary to calculate the lower and upper bound of $\frac{\sigma_A^2}{\sigma_B^2}$. The interval is calculated as follows:

$$\frac{\sigma_A^2}{\sigma_B^2} \in \left[\frac{S_A^2}{S_B^2} F_{1-\frac{\alpha}{2},n_B-1,n_A-1}, \frac{S_A^2}{S_B^2} F_{\frac{\alpha}{2},n_A-1,n_B-1}\right]$$

After writing the appropriate code in the $Q2\ a.R$ file, the bounds of the variances are $[0.9816425, 1.3448907]$ under 95% confidence. Since $1 \in [0.9816425, 1.3448907]$ and both bounds are very close to 1, it is safe to assume that the variances are identical.

Computing the lower and upper bound of the difference of the means of the plastic waste collected weekly by both companies, namely $\mu_A - \mu_B$, under 0.05 significance level, using the following formula:

$$\mu_A - \mu_B \in \left[\bar{X}_A - \bar{X}_B - t_{\frac{\alpha}{2},n_A+n_B-2} S_p \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}, \bar{X}_A - \bar{X}_B + t_{\frac{\alpha}{2},n_A+n_B-2} S_p \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right]$$

$$\text{Where } S_p = \sqrt{\frac{(n_A-1)S_A^2+(n_B-1)S_B^2}{n_A-n_B-2}}$$

the bounds are $[-1592.119, -1577.311]$. This may indicate that company B collects more plastic overall in a week. This is true; however, company B also manages 4 more regions. Doing this analysis for plastic waste collected per week per region, $\mu_A - \mu_B \in [26.34463, 41.15237]$, which indicates that company A collects, on average, more plastic waste per week per region than company B does.

Glass Waste Collected by Companies A and B:

The same steps will be repeated to check the equivalence of the variances:

$\frac{\sigma_A^2}{\sigma_B^2} \in [0.4324050, 0.5924128]$: $1 \notin [0.4324050, 0.5924128]$, so variances are not identical, under 95% confidence, and glass waste collected by company B has a higher variance than that of A. Since variances are not equal, the following formula will be used:

$$\mu_A - \mu_B \in \left[\bar{X}_A - \bar{X}_B - t_{\frac{\alpha}{2},v} \sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}}, \bar{X}_A - \bar{X}_B + t_{\frac{\alpha}{2},v} \sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}}\right]$$

$$\text{Where } v = \frac{\left(\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}\right)^2}{\frac{\left(\frac{S_A^2}{n_A}\right)^2}{n_A+1}+\frac{\left(\frac{S_B^2}{n_B}\right)^2}{n_B+1}}$$

Checking for the bounds of the difference of the weekly average collected waste, which is $[-903.2225, -895.7444]$, which suggests that company B collects more glass weekly. As the previous comparison, the bounds of the difference of the weekly average waste collected per region will be studied, which is $[-12.49378,\ -5.03495]$. This suggests that company B collects more glass waste than company A per region each week, albeit the difference is negligible compared to the average waste collected weekly per region by both companies.

Plastic and Glass Waste Collected by Company A:

Ratio of variances' bounds: $\frac{\sigma_P^2}{\sigma_G^2} \in [5.158448, 7.189092]$. So, variances are not identical, and plastic waste has more variance than collected glass waste.

Since both types of data are generated by the same company, there is no need to check for weekly average per region; weekly average is enough, and the formula used will be the same as the previous part:

$\mu_P - \mu_G \in [6135.677, 6148.235]$ under 0.05 significance level, indicating that company A collects a lot more plastic waste than glass waste each week.

Note: for all the parts above, the normality of data is assumed in order to make statistically sound assumptions.

Q2-b) The city council thinks that the weekly ratio of non-recyclable waste (in the total wast…

To find the expected long-run ratio of non-recycled waste to total waste for each region, grouping the data into appropriate categories is needed. In R-Script, a data frame with regions as headers and weeks as rows is made. The non-recycling ratio is calculated for every region in each week.

Week	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6	Region 7	Region 8	Region 9	Region 10	Re 11
1	0.27044401	0.27887191	0.15357989	0.14957448	0.28709429	0.08239269	0.08077303	0.18087860	0.25198958	0.22716822	0.
2	0.11692914	0.09713110	0.21093995	0.21006170	0.09555116	0.06948060	0.22808018	0.24204623	0.06426560	0.12244863	0.
3	0.24324585	0.14499324	0.29324968	0.11971200	0.15133387	0.06433618	0.19362786	0.09002715	0.26510166	0.13033212	0.
4	0.11163350	0.14784550	0.07471452	0.21801638	0.12751155	0.21280637	0.25031302	0.14609578	0.17223381	0.14095664	0.
5	0.12710020	0.26319004	0.09810138	0.24931125	0.27392239	0.23524528	0.16465355	0.29669386	0.05347568	0.11009888	0.
6	0.14418041	0.05969471	0.11798408	0.19786301	0.26649939	0.22384195	0.13013327	0.31354007	0.05742571	0.18892826	0.
7	0.14694697	0.26367525	0.27734318	0.25731910	0.14428641	0.25230022	0.17081662	0.06091543	0.07616250	0.08395245	0.
8	0.10626857	0.27927476	0.25607735	0.08915010	0.20764242	0.25880695	0.29422222	0.18920742	0.19451889	0.09582395	0.
9	0.22861597	0.10428134	0.16927391	0.21845736	0.25589134	0.08536422	0.17706120	0.07867744	0.17891358	0.04582197	0.
10	0.03679273	0.27573602	0.19288760	0.10739554	0.28476099	0.14911553	0.26498407	0.20643813	0.31328862	0.11384731	0.
11	0.06912907	0.16424563	0.25193015	0.09395232	0.16621185	0.19271410	0.17469777	0.11426736	0.23435565	0.14364716	0.
12	0.14411393	0.13138727	0.16971099	0.23831771	0.23627523	0.13608061	0.13582896	0.29463649	0.15366773	0.23982388	0.
13	0.05261245	0.20508895	0.25944008	0.05479675	0.25923678	0.23180736	0.03624676	0.06996202	0.28153268	0.20887852	0.
14	0.17053257	0.10619998	0.19078655	0.14802387	0.15835557	0.15456854	0.22913353	0.31741922	0.22518676	0.09373074	0.

Figure 8: Weekly non-recycling ratio of total waste for each region.

Then, confidence intervals are calculated for each region on its own, to find the long-run non-recycling ratio. Since the objective is to determine whether the mean of the ratio is below

18% or not, a one-sided, upper-bound confidence interval is used, and the upper bound is calculated, as follows:

$$\mu_R \in \left[0, \bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right]$$

At 95% confidence, it is determined that 32 regions' non-recycling ratio's upper bound is above 18%, thus these regions need awareness-raising campaigns. The only four regions whose ratio's mean's upper bounds were less than 18% are regions 1, 10, 15, and 31.

As for the non-recycling ratio across the whole city, the upper bound is found out to be 0.174, or 17.4%. So, city-wide campaign awareness will not be needed.

Q2-c) Companies state that when the ratio of wastes that are not suitable for the contain…

After calculating the ratio of non-recycled waste to total waste for every week and every waste type for each, it is now easy to find which weeks are lost for every waste type, as shown in Appendix B1 and B2. For both companies, there were no lost weeks for both the glass wastes and aluminum wastes. As a result, there is no need to check whether the lost weeks for glass and aluminum were distributed among both companies equally. For the plastic waste types, however, most of the weeks were lost, at a proportion of 0.8 and about 0.86 for companies A and B respectively.

To find out the difference between the proportion of weeks lost by company A, p_A, and that of B, p_B, confidence intervals will be used, as such:

$$p_A - p_B \in \left[\hat{p}_A - \hat{p}_B - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}, \hat{p}_A - \hat{p}_B + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}\right]$$

Since Z score is used in the calculation, normality of data will be assumed, in order to get meaningful and accurate results.

Now, calculating the confidence interval of the difference between the proportion of lost weeks at 95% confidence, the following interval is found: $p_A - p_B \in [-0.2332, 0.1189]$. Since 0 lies withing the interval, there is not enough evidence to falsify the claim of the municipality,

who claims that the regions were distributed among both companies such that the proportion of lost weeks is similar.

Appendix A

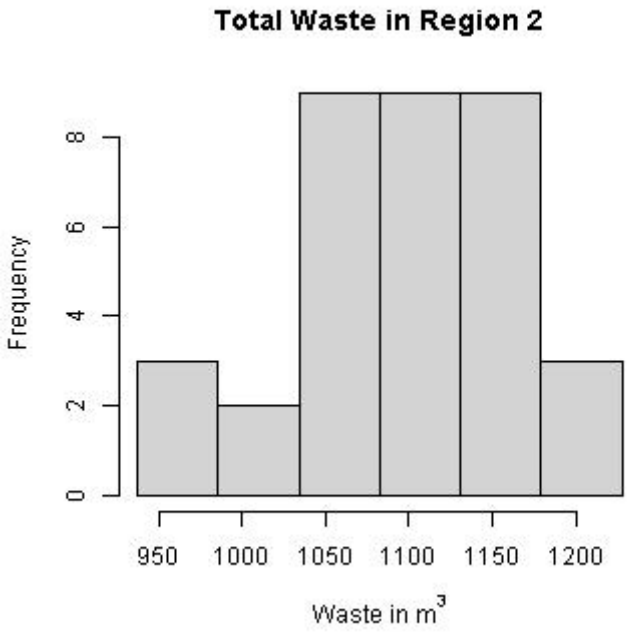


Figure A1: Example of a histogram of weekly total waste that is not normal.

Even though the histogram weekly plastic waste in region 2 is somehow normal, the histogram of total waste is not.

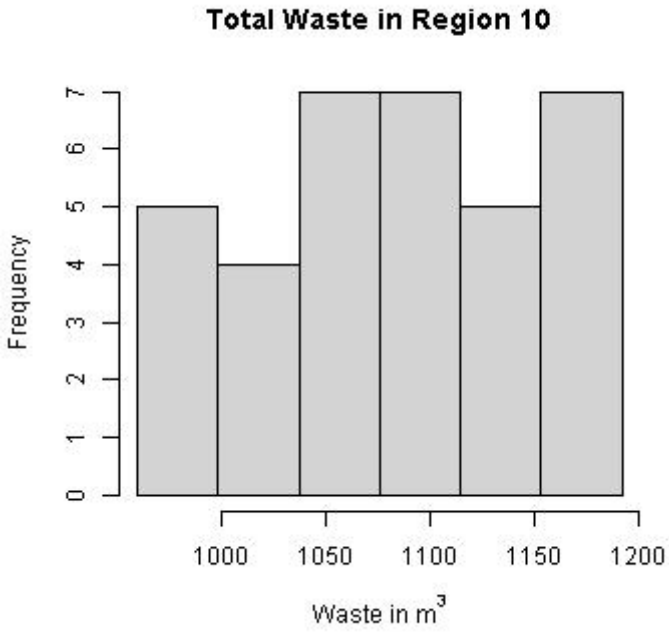


Figure B2: Example of a histogram that is far from normality.

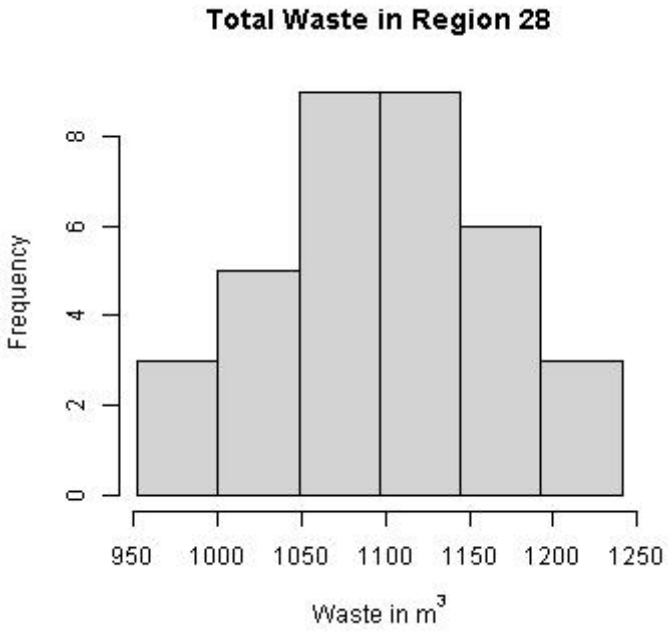


Figure 3: Example of a histogram that is somehow normal.

APPENDIX B

Week	Plastic	Glass	Aluminum
1	TRUE	FALSE	FALSE
2	FALSE	FALSE	FALSE
3	TRUE	FALSE	FALSE
4	TRUE	FALSE	FALSE
5	TRUE	FALSE	FALSE
6	FALSE	FALSE	FALSE
7	TRUE	FALSE	FALSE
8	TRUE	FALSE	FALSE
9	TRUE	FALSE	FALSE
10	TRUE	FALSE	FALSE
11	TRUE	FALSE	FALSE
12	TRUE	FALSE	FALSE
13	TRUE	FALSE	FALSE
14	TRUE	FALSE	FALSE
15	TRUE	FALSE	FALSE
16	FALSE	FALSE	FALSE
17	TRUE	FALSE	FALSE
18	TRUE	FALSE	FALSE
19	TRUE	FALSE	FALSE
20	TRUE	FALSE	FALSE
21	TRUE	FALSE	FALSE
22	TRUE	FALSE	FALSE
23	TRUE	FALSE	FALSE
24	TRUE	FALSE	FALSE
25	FALSE	FALSE	FALSE
26	FALSE	FALSE	FALSE
27	TRUE	FALSE	FALSE
28	FALSE	FALSE	FALSE
29	TRUE	FALSE	FALSE
30	FALSE	FALSE	FALSE
31	TRUE	FALSE	FALSE
32	TRUE	FALSE	FALSE
33	TRUE	FALSE	FALSE
34	TRUE	FALSE	FALSE
35	TRUE	FALSE	FALSE

Week	Plastic	Glass	Aluminum
1	TRUE	FALSE	FALSE
2	TRUE	FALSE	FALSE
3	TRUE	FALSE	FALSE
4	FALSE	FALSE	FALSE
5	TRUE	FALSE	FALSE
6	TRUE	FALSE	FALSE
7	TRUE	FALSE	FALSE
8	TRUE	FALSE	FALSE
9	TRUE	FALSE	FALSE
10	TRUE	FALSE	FALSE
11	FALSE	FALSE	FALSE
12	TRUE	FALSE	FALSE
13	FALSE	FALSE	FALSE
14	TRUE	FALSE	FALSE
15	TRUE	FALSE	FALSE
16	TRUE	FALSE	FALSE
17	TRUE	FALSE	FALSE
18	TRUE	FALSE	FALSE
19	TRUE	FALSE	FALSE
20	TRUE	FALSE	FALSE
21	FALSE	FALSE	FALSE
22	TRUE	FALSE	FALSE
23	TRUE	FALSE	FALSE
24	TRUE	FALSE	FALSE
25	TRUE	FALSE	FALSE
26	TRUE	FALSE	FALSE
27	TRUE	FALSE	FALSE
28	TRUE	FALSE	FALSE
29	TRUE	FALSE	FALSE
30	TRUE	FALSE	FALSE
31	TRUE	FALSE	FALSE
32	TRUE	FALSE	FALSE
33	TRUE	FALSE	FALSE
34	TRUE	FALSE	FALSE
35	FALSE	FALSE	FALSE

Figures B1 and B2: Weeks lost for each data type, for companies A and B respectively.