# Objective 2

We ran a two-way ANOVA on the data to satisfy the second objective. The business requirements drove the selection of the explanatory variables. The business wants to know what time of day and on which server new reports should be added based upon current report run times as expressed by the ReportDeliveryTime response variable.

## Explanatory and Response variable creation

The two explanatory variables require some explanation. The “server” variable required us to remove the data associated with testing and development servers. The testing and development report run times have very different characteristics than production data. Thus, all data associated with servers “SQLODR1” and “SQLODR5” were removed.

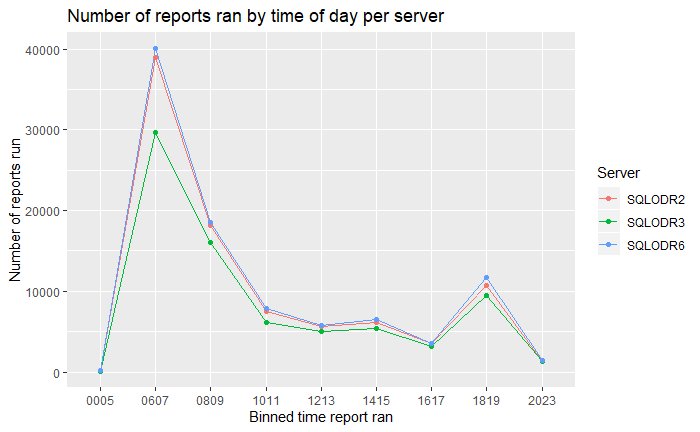
“HourBinned” variable indicates the time of day the report ran. Several things should be noted about it. A maintenance window exists every night between the hours of midnight and six AM. Only a few reports run during this maintenance window. Likewise, relatively few reports run during the late evening hours between eight PM and midnight. We binned these two time periods separately. The remaining time during the days was binned in two-hour blocks. The summary tables below give counts and means of the various bins.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| server  <fctr> | N  <dbl> | Mean  <dbl> | SD  <dbl> | SE  <dbl> |
| SQLODR2 | 92387 | 2262.705 | 7028.219 | 0.07607368 |
| SQLODR3 | 76354 | 1976.964 | 6562.348 | 0.08594636 |
| SQLODR6 | 95402 | 2168.218 | 6913.200 | 0.07246389 |

The table above shows that server “SQLODR3” runs fewer reports than the other two servers and has differences in mean report run time and standard deviation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| HourBinned  <fctr> | N  <dbl> | Mean  <dbl> | SD  <dbl> | SE  <dbl> |
| 0005 | 384 | 69285.472 | 13610.750 | 35.44466176 |
| 0607 | 108726 | 2021.303 | 4943.980 | 0.04547192 |
| 0809 | 52831 | 1624.391 | 5744.695 | 0.10873720 |
| 1011 | 21445 | 3045.266 | 7553.906 | 0.35224555 |
| 1213 | 16435 | 2525.486 | 9787.416 | 0.59552270 |
| 1415 | 18022 | 1483.042 | 5885.803 | 0.32658992 |
| 1617 | 10284 | 3169.675 | 8681.578 | 0.84418305 |
| 1819 | 31846 | 1358.025 | 5025.376 | 0.15780243 |
| 2023 | 4170 | 6059.903 | 16035.427 | 3.84542610 |

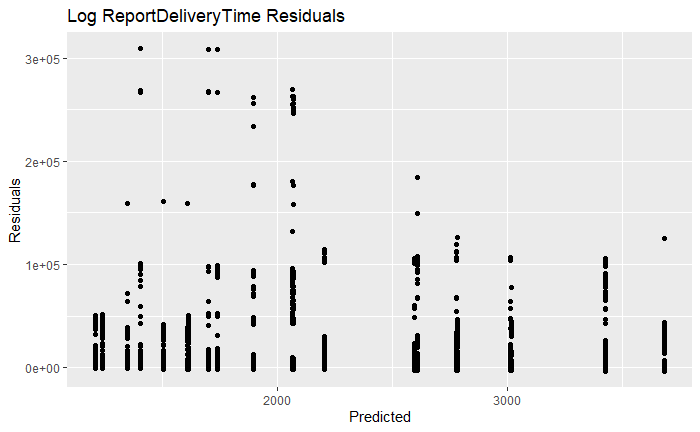
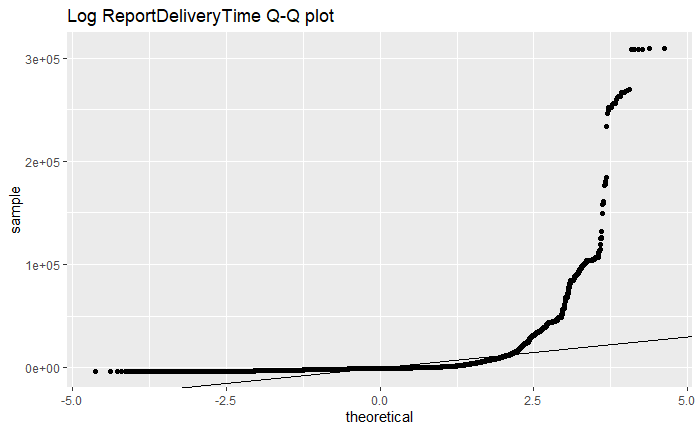
Looking at the data above for the BinnedHour variable the late evening hours and the overnight maintenance window bins display very different standard deviations than the rest of the bins. Also, of note the sample count size differs for those two bins. Finally reports should not be added to the overnight bin that includes the maintenance window.

 The graph show that each server tracks the others in terms of reports run at a particular hour. A simple interpretation of the results benefits the business directly. The first analysis of the data included an unlogged response variable. The bins for overnight and late evening (“0005”,”2023”)are excluded from the analysis.

**ANOVA Type III table for unlogged response variable**

|  |  |
| --- | --- |
| Term | p-value |
| (Intercept) | < 2.2e-16 |
| Server | 0.0001712 |
| HourBinned | < 2.2e-16 |
| Server:HourBinned | < 2.2e-16 |

The type 3 ANOVA table indicates that both terms as well as the interaction term are statistically significant with very low p-values.

The residual plot appears to demonstrate a cone shape with residuals beneath 2000 being noticeably more spread out than those above 3000. The Q-Q plot displays dramatic differences from normality. A log transform of the response variable is necessary.

The summary tables for the logged response variable are displayed below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| server  <fctr> | N  <dbl> | Mean  <dbl> | SD  <dbl> | SE  <dbl> |
| SQLODR2 | 90872 | 6.150445 | 2.098771 | 2.309590e-05 |
| SQLODR3 | 74945 | 5.977121 | 2.099904 | 2.801926e-05 |
| SQLODR6 | 93772 | 6.072446 | 2.135042 | 2.276844e-05 |

The mean and standard deviation of the logged ReportDeliveryTime track extremely close now with respect to the server variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| HourBinned  <fctr> | N  <dbl> | Mean  <dbl> | SD  <dbl> | SE  <dbl> |
| 0607 | 108726 | 7.134787 | 1.200840 | 1.104465e-05 |
| 0809 | 52831 | 5.645390 | 2.024127 | 3.831325e-05 |
| 1011 | 21445 | 5.060337 | 2.841504 | 1.325019e-04 |
| 1213 | 16435 | 4.503144 | 2.569200 | 1.563249e-04 |
| 1415 | 18022 | 4.776321 | 2.349070 | 1.303446e-04 |
| 1617 | 10284 | 5.155736 | 2.742066 | 2.666342e-04 |
| 1819 | 31846 | 5.673145 | 1.672099 | 5.250578e-05 |

The HourBinned variable does display some differences in standard deviation. This issue is examined below. Please note that the over night and late evening hours were dropped from the model.

The difference in standard deviation between the various groups is concerning. On-line research yielded the following article <https://link.springer.com/article/10.3758/s13428-017-0918-2?shared-article-renderer#Tab6> . The article indicates that a useful measure of unequal variance is the *Variance ratio*. This ratio is simply the largest group variance divided by the smallest group variance. Looking at the summary statistics table for hours binned we find that:

The correlation between the sample size and variance is -0.81. The coefficient of sample size variation (Δn) was calculated to be

According to the article’s conclusions with the parameters as calculated highlighted below (Table 10 of the article) the Type I error ratio will be liberal. However, most p-values reported by the two-way ANOVA are very small. The effects of unequal sample size and unequal variance within the sample groups should only affect relatively few term pairings. Assume p-values can be as much as four times larger than the given value.

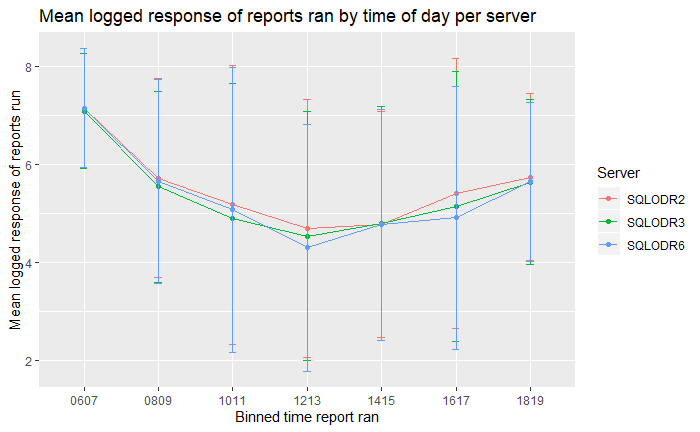
|  |  |  |  |
| --- | --- | --- | --- |
| Variance ratio | Pairing | Coefficient of sample size variation | Type I error rate |
| 1.6, 1.7, 1.8 | + or −1 | 0.50 | Liberal |
| 2 | −1 | 0.33; 0.50 | Liberal |
| 3 | 1 | 0.50 | Conservative |
|  | −.50 | 0.50 | Liberal |
|  | −1 | 0.33; 0.50 | Liberal |
| 5 & 9 | 1 | 0.50 | Conservative |
|  | −.50 | 0.33; 0.50 | Liberal |
|  | −1 | 0.16; 0.33; 0.50 | Liberal |

**ANOVA Type III table for logged response variable**

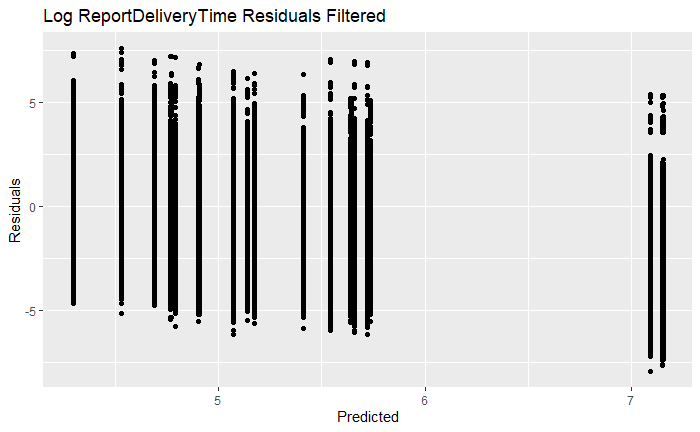
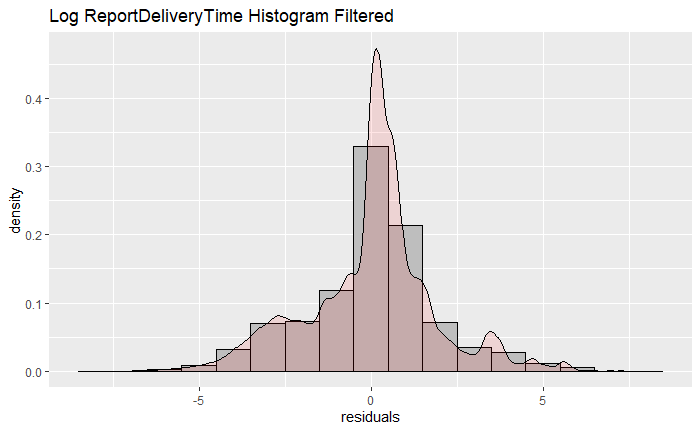
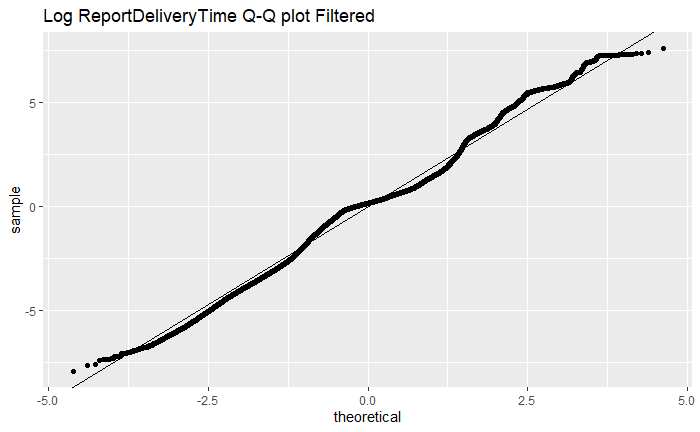
|  |  |
| --- | --- |
| Term | p-value |
| (Intercept) | < 2.2e-16 |
| Server | 7.631e-06 |
| HourBinned | < 2.2e-16 |
| Server:HourBinned | < 2.2e-16 |

The type III ANOVA table again shows that both the interaction term as well as the individual terms are significant.

The following graph shows that the mean log of the delivery time varies by the time of day and has a different standard deviation during times of high loads.



The following plots were created to further check for normality

Residuals, Q-Q plot, and the Histogram all support the assumption of normality.

Finally, a pairwise comparison based upon server and HourBinned was run. Remember to adjust p-value by a factor of four due to the unequal variance and sample sizes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Servers | diff | lwr | upr | p adj |
| SQLODR3-SQLODR2 | -0.17332340 | -0.19508443 | -0.15156236 | 0.000000e+00 |
| SQLODR6-SQLODR2 | -0.07799805 | -0.09852703 | -0.05746906 | 2.131628e-14 |
| SQLODR6-SQLODR3 | 0.09532535 | 0.07371693 | 0.11693377 | 2.875478e-14 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| HourBinned | diff | lwr | upr | p adj |
| 0809-0607 | -1.48524162 | -1.51466386 | -1.45581938 | 0.0000000000 |
| 1011-0607 | -2.07262721 | -2.11407977 | -2.03117465 | 0.0000000000 |
| 1213-0607 | -2.62750188 | -2.67393276 | -2.58107100 | 0.0000000000 |
| 1415-0607 | -2.35477484 | -2.39939451 | -2.31015518 | 0.0000000000 |
| 1617-0607 | -1.97481318 | -2.03204901 | -1.91757736 | 0.0000000000 |
| 1819-0607 | -1.45780580 | -1.49315497 | -1.42245664 | 0.0000000000 |
| 1011-0809 | -0.58738559 | -0.63230570 | -0.54246549 | 0.0000000000 |
| 1213-0809 | -1.14226026 | -1.19181153 | -1.09270900 | 0.0000000000 |
| 1415-0809 | -0.86953322 | -0.91739151 | -0.82167493 | 0.0000000000 |
| 1617-0809 | -0.48957157 | -0.54936656 | -0.42977658 | 0.0000000000 |
| 1819-0809 | 0.02743581 | -0.01192235 | 0.06679398 | 0.3797535230 |
| 1213-1011 | -0.55487467 | -0.61238966 | -0.49735968 | 0.0000000000 |
| 1415-1011 | -0.28214763 | -0.33821065 | -0.22608461 | 0.0000000000 |
| 1617-1011 | 0.09781403 | 0.03127015 | 0.16435790 | 0.0002948882 |
| 1819-1011 | 0.61482141 | 0.56581415 | 0.66382867 | 0.0000000000 |
| 1415-1213 | 0.27272704 | 0.21288914 | 0.33256494 | 0.0000000000 |
| 1617-1213 | 0.65268869 | 0.58293485 | 0.72244254 | 0.0000000000 |
| 1819-1213 | 1.16969607 | 1.11641171 | 1.22298044 | 0.0000000000 |
| 1617-1415 | 0.37996166 | 0.31140010 | 0.44852322 | 0.0000000000 |
| 1819-1415 | 0.89696904 | 0.84525529 | 0.94868278 | 0.0000000000 |
| 1819-1617 | 0.51700738 | 0.45408412 | 0.57993064 | 0.0000000000 |

**Conclusion and discussion**

The business asked that we determine on what server and at what time of day additional reports should run. Due to the residual and Q-Q plot of the untransformed data it was found to be necessary to perform a log transform of the response data. The data responded well to the transformation displaying all the characteristics of normalized data.

Using the table “Mean logged response of reports ran by time of day per server” above we can make the following recommendations.

* The number of reports run on server “SQLODR03” is about 20 % less than the other two servers. However, at most it has only 18% faster mean report delivery time(95% CI[5.3% - 32%]) this is between server “SQLODR03” and “SQLODR6” for the 10 AM through noon time period
* The time period 6 AM to 8 AM has the highest mean report delivery time, followed by 8 AM to 10 AM and 6 PM to 8 pm. Every effort should be made to not schedule new reports during these time frames.
* The time period noon to 2 PM has the lowest mean report delivery time, thus if possible new reports should be added there. The relative loading of the servers is as follows. Server “SQLODR02” has a 48% greater mean report delivery time (95% CI[31%, 68%]) than server”SQLODR06” for that time frame. While server “SQLODR3” has a 27% greater mean report delivery time (95% CI[11% , 44%]) than server ”SQLODR06”. Server ”SQLODR06” is the lightest loaded server as measured by mean report delivery time during the noon to 2PM time frame.
* The very late evening hours 8PM to midnight remain an option for scheduling reports. Care must be taken that any reports scheduled during that time do not run past midnight into the maintenance window.