# 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.

## 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.

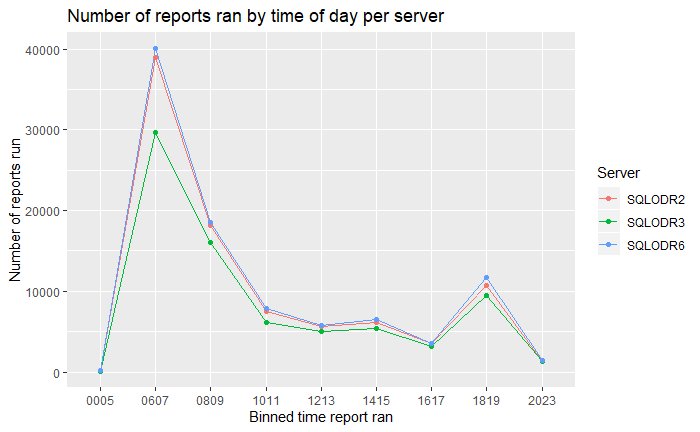
“HourBinned” variable indicates the time of day the report ran. Several things should be noted about it. A maintenance window exists every night between the hours of midnight and six AM. Only a few reports run during this maintenance window. Likewise, relatively few reports run during the late evening 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 |

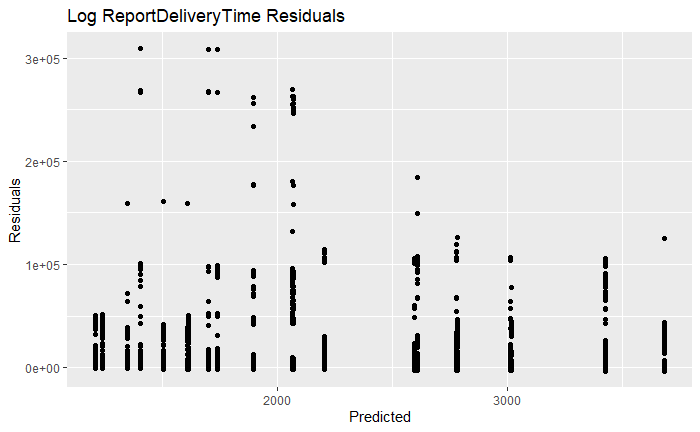
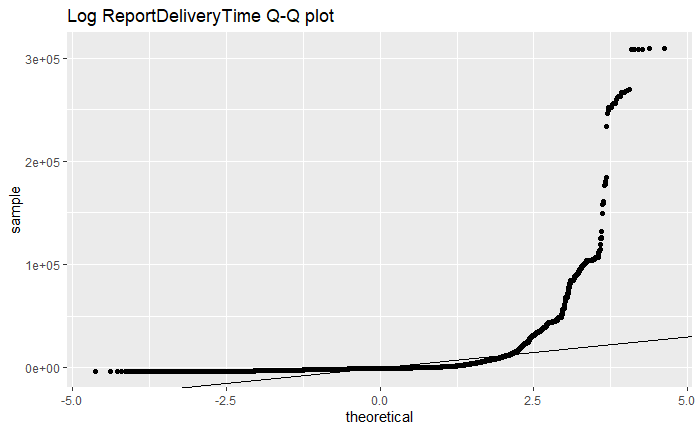
Looking at the data above for the BinnedHour variable the late evening hours and the overnight maintenance window bins display very different standard deviations than the rest of the bins. Also, of note the sample count size differs for those two bins. Finally reports should not be added to the overnight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

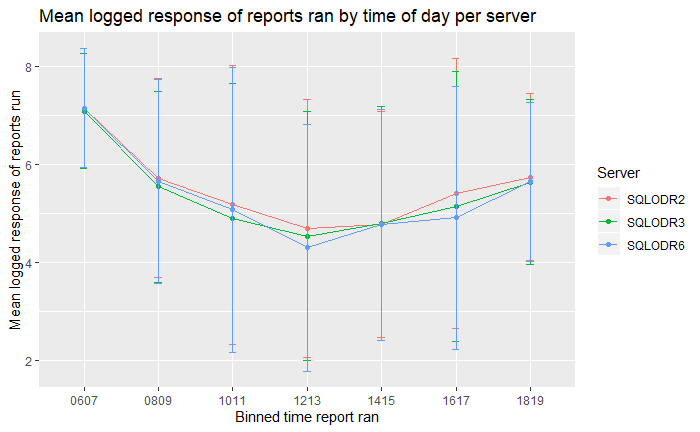
The HourBinned variable does display some differences in standard deviation. However, the extremely high sample sizes should protect us from invalid model response. Please note that the over night and late evening hours were dropped from the model.

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

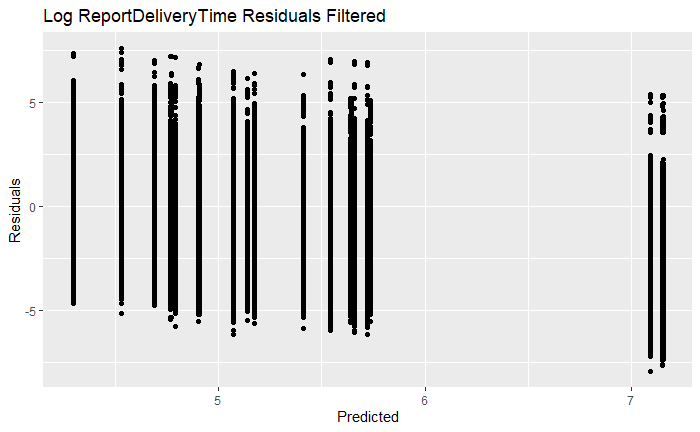
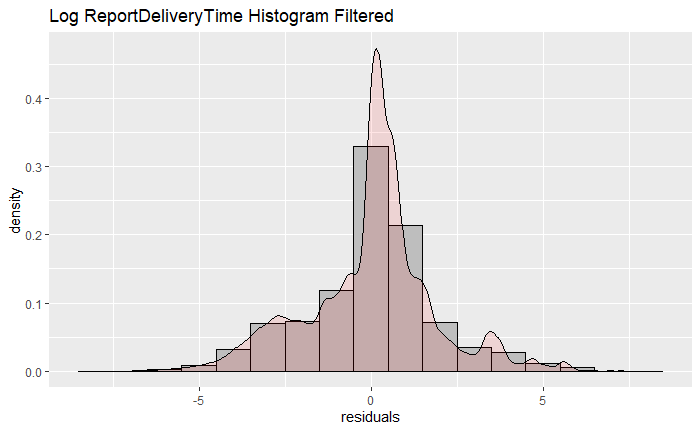
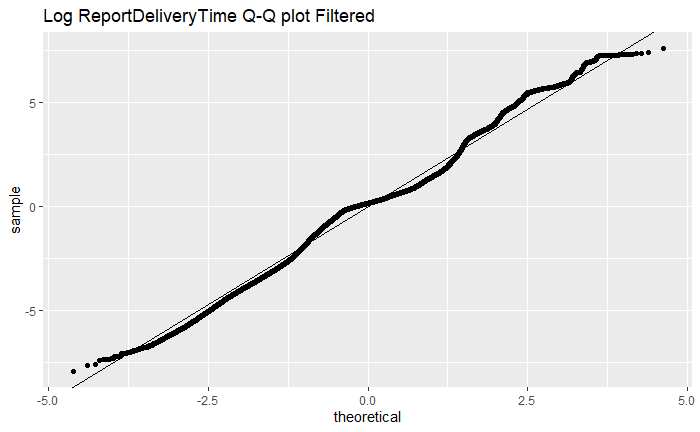
|  |  |
| --- | --- |
| Term | p-value |
| (Intercept) | < 2.2e-16 |
| Server | 7.631e-06 |
| HourBinned | < 2.2e-16 |
| Server:HourBinned | < 2.2e-16 |

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.