TimelyCare Assesment

Tracy Reuter

Question 1 (SQL code)

We're noticing a data quality issue and want you to investigate. What's the total number of visits that cannot occur because the provider was double-booked at the same time? What is the total count that can not continue because the provider didn't have a valid license in the member's state?

```
# -- written in TSQL but translatable to MySQL
# -- 1. identify double-booked appointments
# select
# scheduled_date_time,
# provider_id,
# count(distinct(member_id)) as n_patients
# from DB.schema.visit_table
# group by
# scheduled_date_time,
# provider_id
# -- same provider, same time, > 1 patient
# having count(distinct(member_id)) >1;
#
# -- 2. identify invalid licenses
# select
# vis.provider_id,
# vis.scheduled_date_time
# from DB.schema.visit_table as vis
# left join DB.schema.member_table as mem on vis.member_id = mem.id
# left join DB.schema.provider_table as prov on vis.provider_id = prov.id
# where
# -- if provider license state does not match visit state, the license is invalid
# prov.License not like '%' + mem.visit_state + '%'
# -- assuming service_line HC means license valid in all states (* notation)
# and vis.service_line not in ('HC');
```

Question 2 (R code)

We're noticing a data quality issue and want you to investigate. What's the total number of visits that cannot occur because the provider was double-booked at the same time? What is the total count that can not continue because the provider didn't have a valid license in the member's state?

- I assumed "double-booked" means that 1 provider had 2 patients scheduled for the same time (i.e. not counting partial overlaps).
- I assumed "valid license" means that the member state matches provider license state *and* service_line matches license type *and* license active is "true" *or* service_line is "HC" (which might indicate MD or other license that is valid in all states).

Table 1: These visits were double-booked.

provider_id	scheduled_date_time
1	2023-01-15 20:00:00

Table 2: These visits had an invalid license.

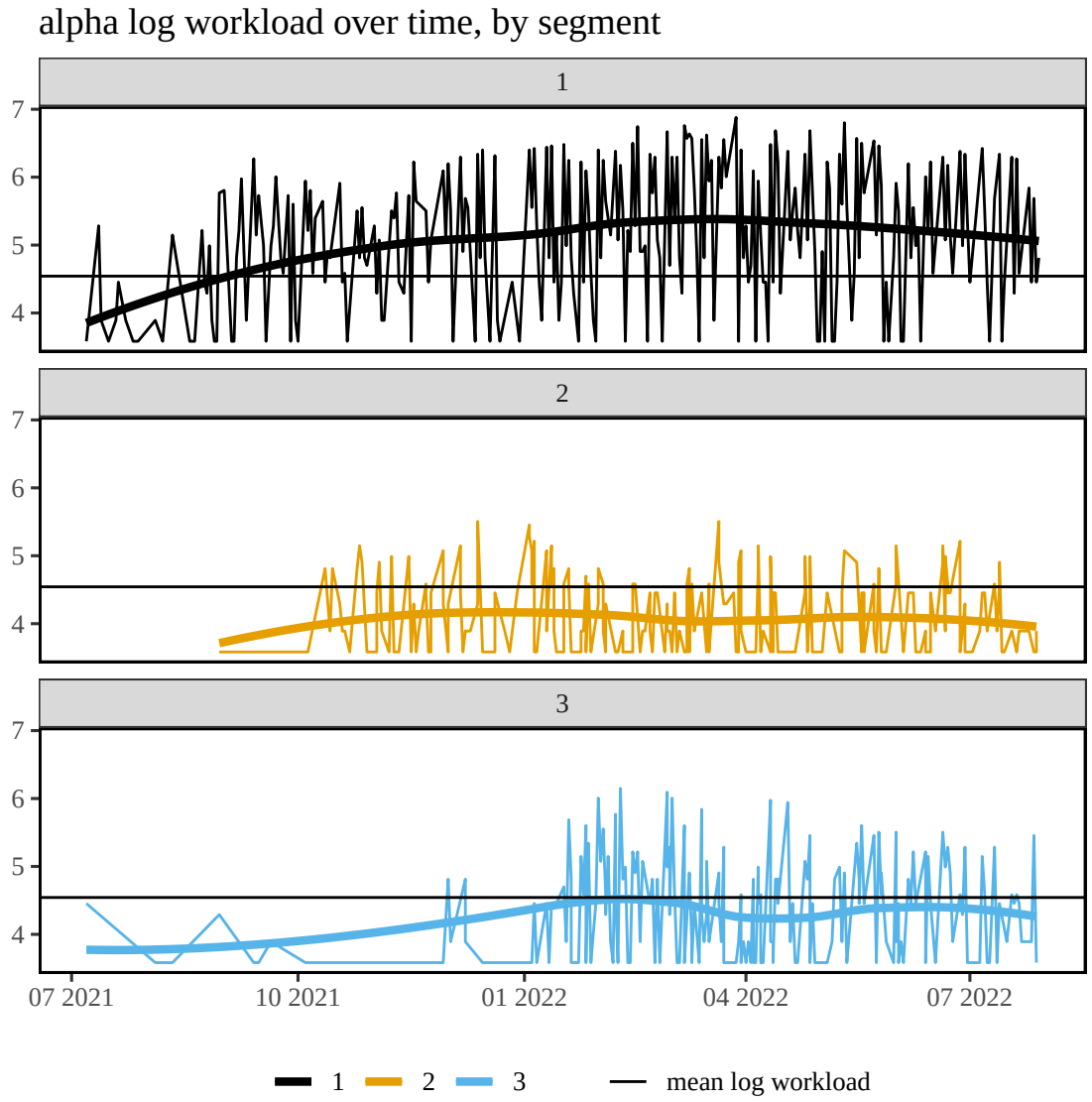
provider_id	scheduled_date_time
2	2023-01-15 20:00:00
5	2023-01-12 19:47:07

Question 3A

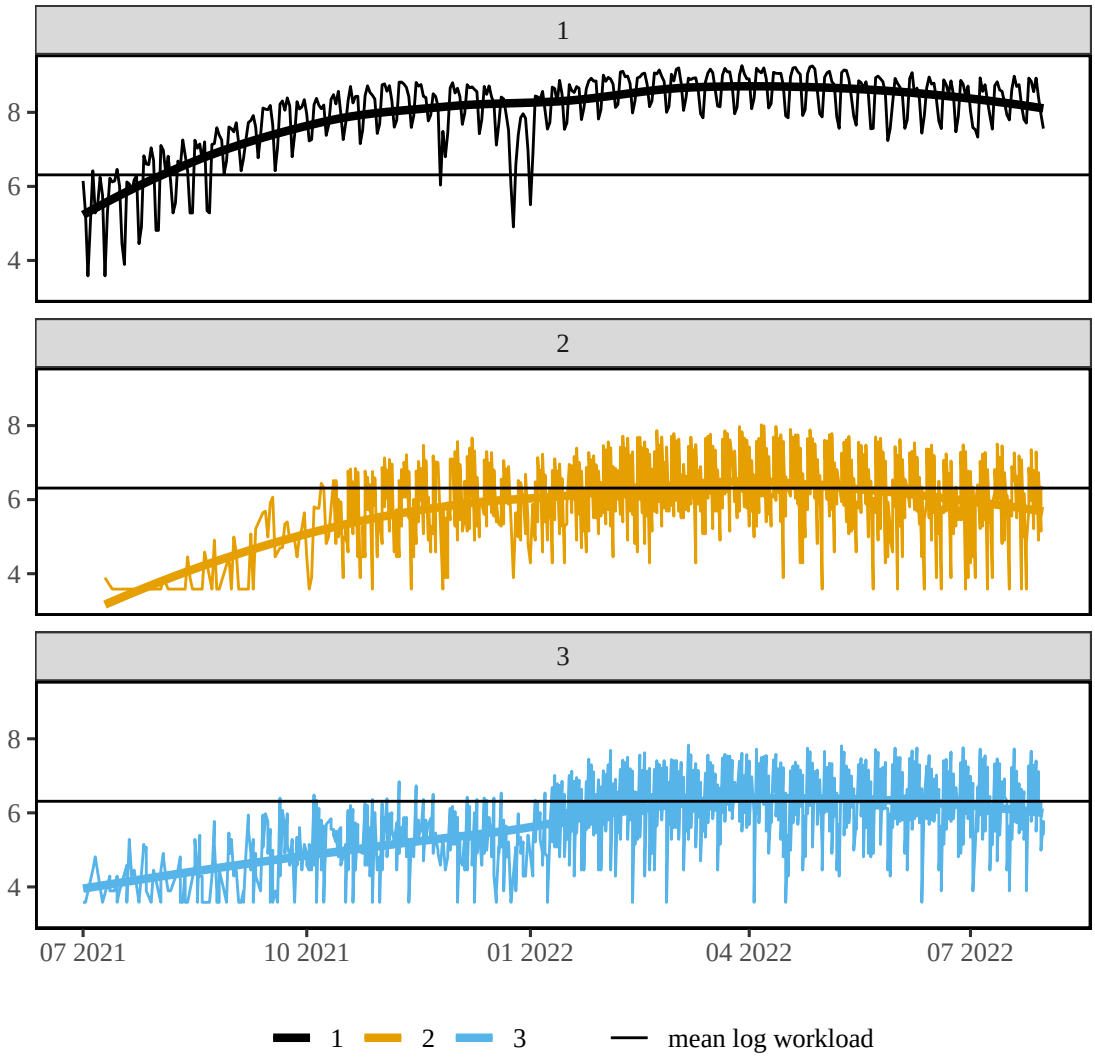
We are interested in using a forecast of the target variable of "workload" to determine how to best staff for a week. With that in mind, please forecast the "workload" by all forecastable "type" for the next 60 days and provide the accuracy measurements you used to determine the best approach for each.

1. Visualize time series by type and segment

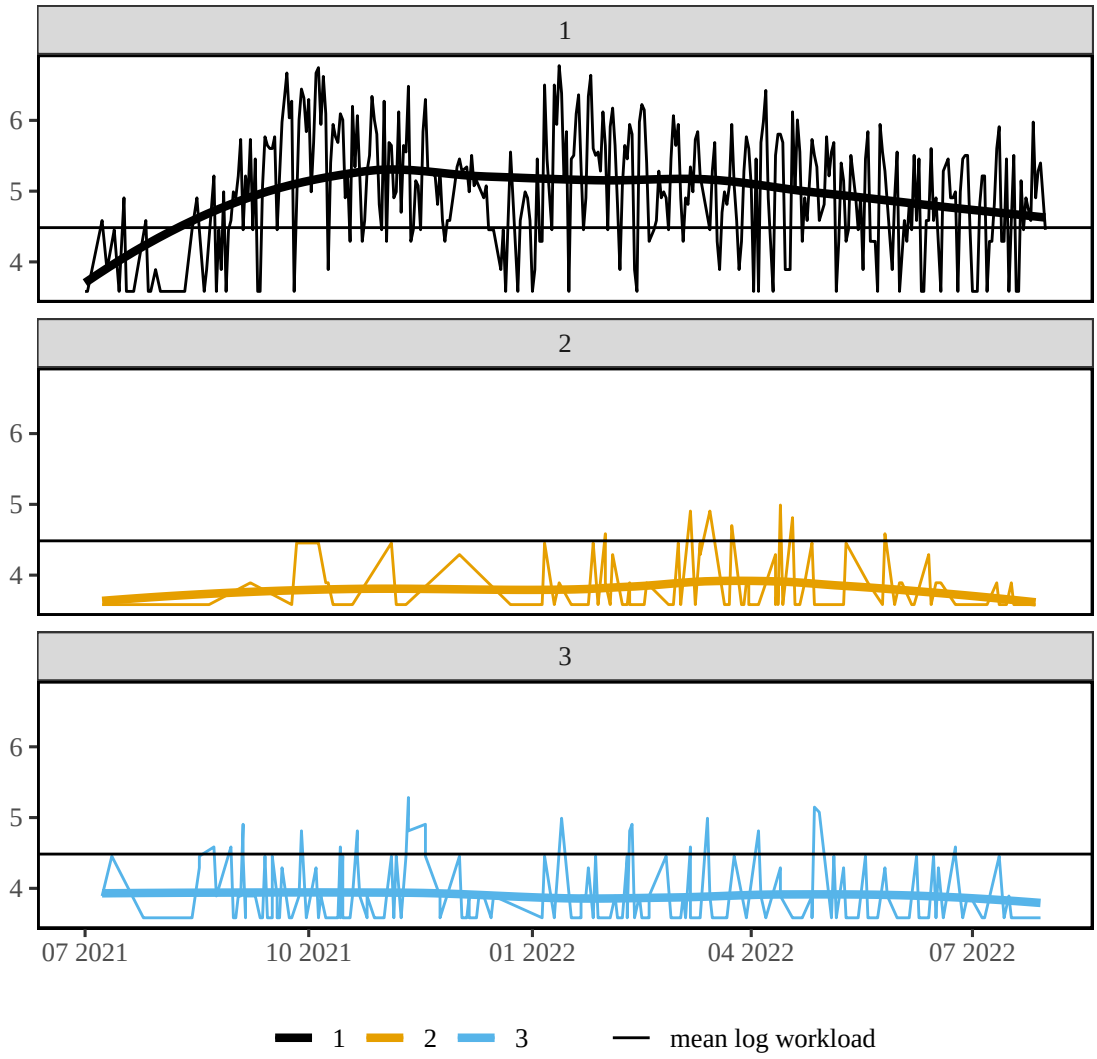
Thin lines display log-transformed workload per date. (Log transformation is detailed in the below code.) Thick lines display a local regression (loess) which helps to quickly visualize trends over time by fitting a smooth curve through the data points. Workload changes differently over time by type (alpha, beta, charlie, delta) and by segment (1, 2, 3).



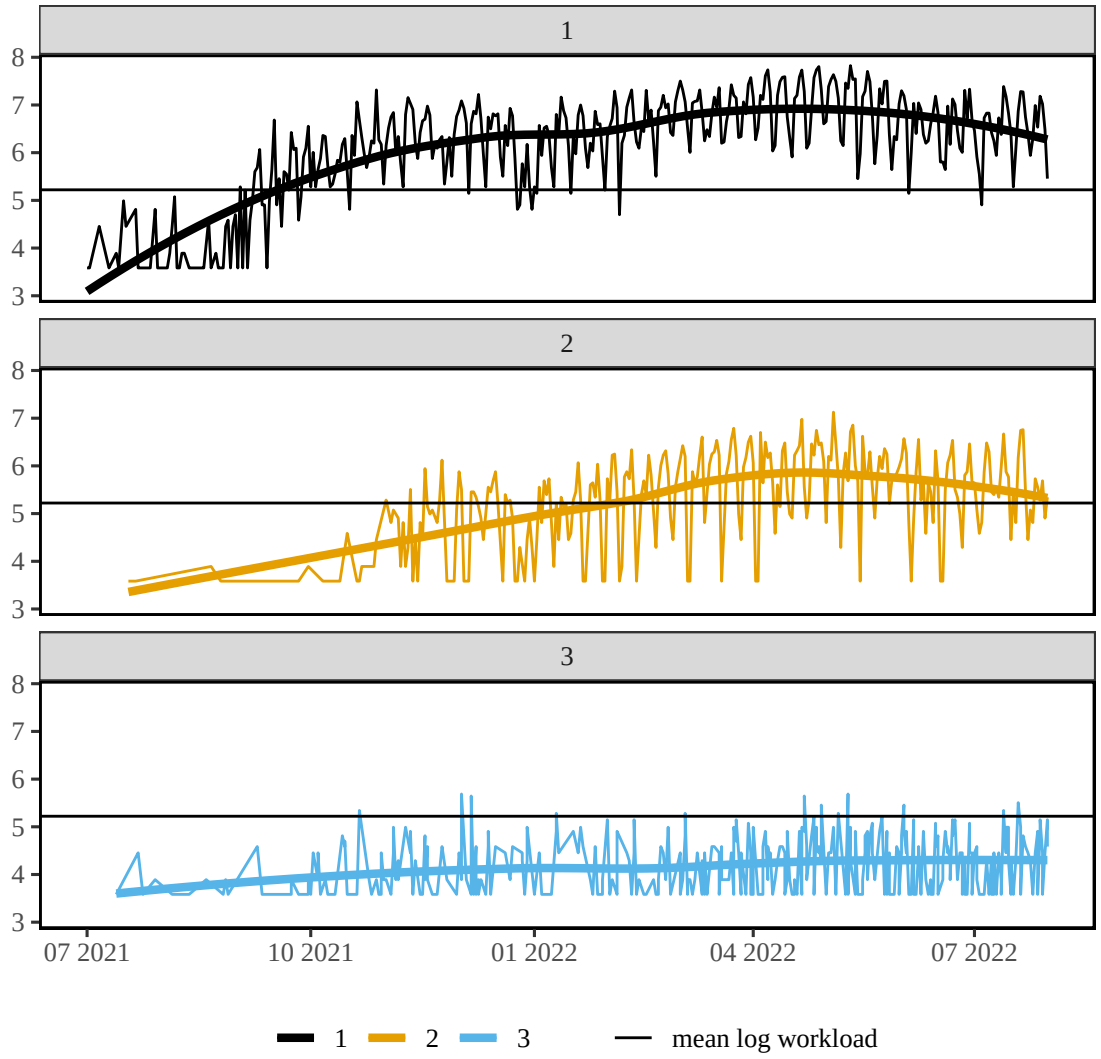
beta log workload over time, by segment



charlie log workload over time, by segment



delta log workload over time, by segment



2. Build time series models and evaluate accuracy

To forecast future workload, I used ARIMA (auto-regressive incremental moving average) models. To evaluate model accuracy, I used BIC, visualized residuals, and used MAPE (mean absolute percentage error) as a final summary statistic. **Detailed explanations are included throughout the below code.** Given that workload varies significantly by segment, I forecasted workload for each type and each segment independently, building 12 total models.

Table 3 summarizes model results for each type and segment. The train_MAPE value indicates how well the model fit the training data, whereas the validation_MAPE value indicates how well the model fit previously-unseen test data. Smaller MAPE values mean better model accuracy. **Importantly, seeing validation_MAPE on par with train_MAPE means that the model was not over-fit to the training data. Rather, it extended well to new data too.** The forecasted_log_workload gives the mean forecasted log workload for the coming 60 days. (The log value helps to compare with the above visuals.) Finally, the forecasted_workload gives the forecast in the original scale.

Table 3: Model Summary Statistics

type	segment	train_MAPE	validation_MAPE	forecasted_log_workload	forecasted_workload
alpha	1	16	15	5	173
alpha	2	12	11	4	60
alpha	3	15	13	4	76
beta	1	5	3	8	4033
beta	2	10	13	6	412
beta	3	11	10	6	457
charlie	1	13	14	5	137
charlie	2	8	6	4	46
charlie	3	9	8	4	50
delta	1	7	6	6	603
delta	2	11	10	5	243
delta	3	11	12	4	76

Question 3B

We are interested in understanding the impact of "segment" on the "workload" forecast. How would you explain the impact the segment has on the overall forecast of "workload"? What "segment" has the most impact on the forecasting of the "workload"?

Based on the above data visualization and model results, **segment 1** has the most impact for forecasting workload. This segment has rapid growth over time, and model results suggest future workload demand will continue.

Question 3C

Explain any seasonality you have discovered in the data.

1. As I developed the modeling function, I used a Ljung-Box test and ACF (auto-correlation function) plots to assess seasonality. The following time series had significant seasonality prior to modeling:

- type alpha, segment 2
- type beta segment 1, 2, and 3
- type charlie, segment 1
- type delta segment 1 and 2

2. The modeling function takes seasonality into account. If the time series had significant seasonality effects, then I directed the ARIMA model selection process to evaluate only seasonal models. Then, selecting the optimal differencing parameter (d) is designed to account for seasonality.

3. Having multiple years of data would help to better understand seasonality. Specifically, there is a decrease in workload around January 2022. Assessing data for January 2023 would help to determine whether that was a one-off spike or an annual pattern.