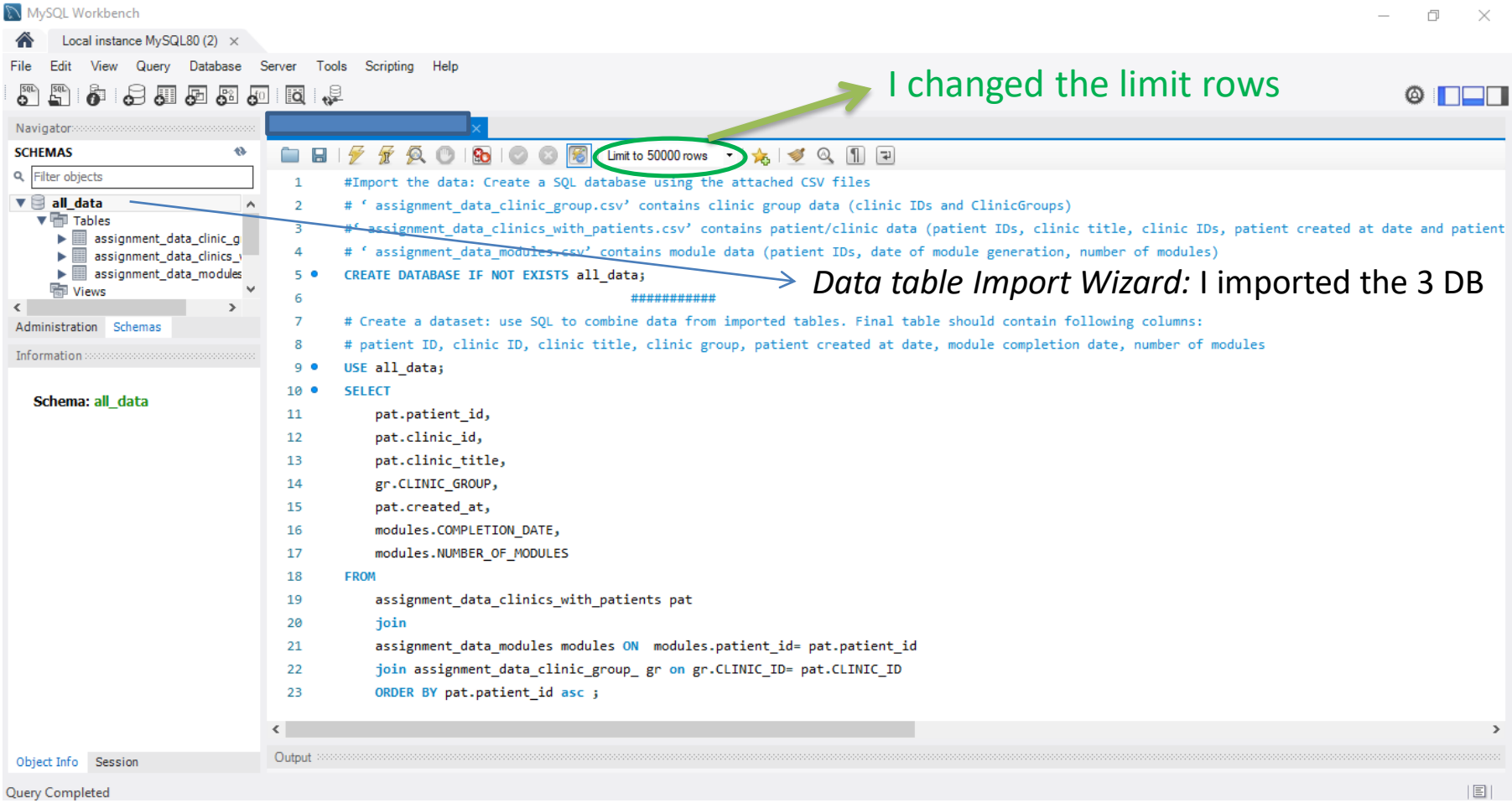
# Elena Peña

SQL and Tableau Assignment

# SQL(1): Code

Difficulties: *Import wizard* couldn't read well the csv-files. There were problems with the encoding, so I went back to Excel reader and inside of "Save as": "Excel options" I changed the file type and "save files in web service". It did work!!



# SQL(2): results

MySQL Workbench

Local instance MySQL80 (2)

File Edit View Query Database Server Tools Scripting Help

Navigator

SCHEMAS

Filter objects

- all_data
  - Tables
    - assignment_data_clinic_group_
    - assignment_data_clinics_with_patients
    - assignment_data_modules
  - Views
  - Stored Procedures

Administration Schemas

Information

Schema: all_data

Object Info Session

Limit to 50000 rows

```
17        modules.NUMBER_OF_MODULES
18    FROM
19        assignment_data_clinics_with_patients pat
20        join
21        assignment_data_modules modules ON  modules.patient_id= pat.patient_id
22        join assignment_data_clinic_group_ gr on gr.CLINIC_ID= pat.CLINIC_ID
23        ORDER BY pat.patient_id asc ;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| patient_id | clinic_id | clinic_title | CLINIC_GROUP | created_at | COMPLETION_DATE | NUMBER_OF_MODULES |
|---|---|---|---|---|---|---|
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 18/10/2020 | 1 |
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 19/10/2020 | 1 |
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 24/10/2020 | 1 |
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 25/10/2020 | 1 |
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 28/10/2020 | 2 |
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 04/11/2020 | 2 |
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 11/11/2020 | 2 |
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 18/11/2020 | 2 |
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 25/11/2020 | 2 |
| 3764 | abc9 | Rehab on the go | Re-Freshed | 04/06/2018 | 02/12/2020 | 2 |

Result Grid Form Editor Field Types

Result 2 Read Only

Output

Action Output

| # | Time | Action | Message | Duration / Fetch |
|---|---|---|---|---|
| 27 | 06:55:57 | USE all_data | 0 row(s) affected | 0.000 sec |
| 28 | 06:55:57 | SELECT pat.patient_id, pat.clinic_id, pat.clinic_title, gr.CLINIC_... | 25798 row(s) returned | 0.312 sec / 0.079 sec |

Query Completed

# SQL(3): Data Exploration

MySQL Workbench

Local instance MySQL80 (2)

File Edit View Query Database Server Tools Scripting Help

Navigator

SCHEMAS

Filter objects

- all_data
  - Tables
    - assignment_data_clinic_group_
    - assignment_data_clinics_with_patients
    - assignment_data_modules
  - Views
  - Stored Procedures

Administration Schemas

Information

No object selected

Object Info Session

Limit to 50000 rows

```
41          assignment_data_modules modules
42        order by  modules.patient_id  ;
43    # more than one row per patient_id
44
45    SELECT
46    (select count(Distinct (pat.patient_id)) from assignment_data_clinics_with_patients pat) as count_distinct_patients_PatientsDB
47     ,
48     (select count(Distinct (modules.patient_id)) from  assignment_data_modules modules) as count_distinct_patients_ModulesDB,
49     (select count(Distinct (pat.patient_id)) from assignment_data_clinics_with_patients pat)
50     -
51     (select count(Distinct (modules.patient_id)) from  assignment_data_modules modules) as Difference_btw_DBs;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| count_distinct_patients_PatientsDB | count_distinct_patients_ModulesDB | Difference_btw_DBs |
|---|---|---|
| 18596 | 2290 | 16306 |

I'm going to analyze 25798 data rows from 2290 different patients

Result Grid

Form Editor

Result 35

Read Only

Output

Action Output

| # | Time | Action | Message | Duration / Fetch |
|---|---|---|---|---|
| 38 | 11:05:19 | SELECT (select count(Distinct (pat.patient_id)) from assignment_data_clinics_with... | 1 row(s) returned | 0.125 sec / 0.000 sec |

# SQL(5): Data Exploration
# Missing values

```
# patient ID, clinic ID, clinic title, clinic group, patient created at date, module completion date, number of modules
USE all_data;
SELECT
    pat.patient_id, pat.clinic_id,
    pat.clinic_title,  gr.CLINIC_GROUP,  pat.created_at,
    modules.COMPLETION_DATE,
    modules.NUMBER_OF_MODULES
FROM
    assignment_data_clinics_with_patients pat
    join
    assignment_data_modules modules ON  modules.patient_id= pat.patient_id
    join assignment_data_clinic_group_ gr on gr.CLINIC_ID= pat.CLINIC_ID
    is null
    ORDER BY pat.patient_id asc ;
```

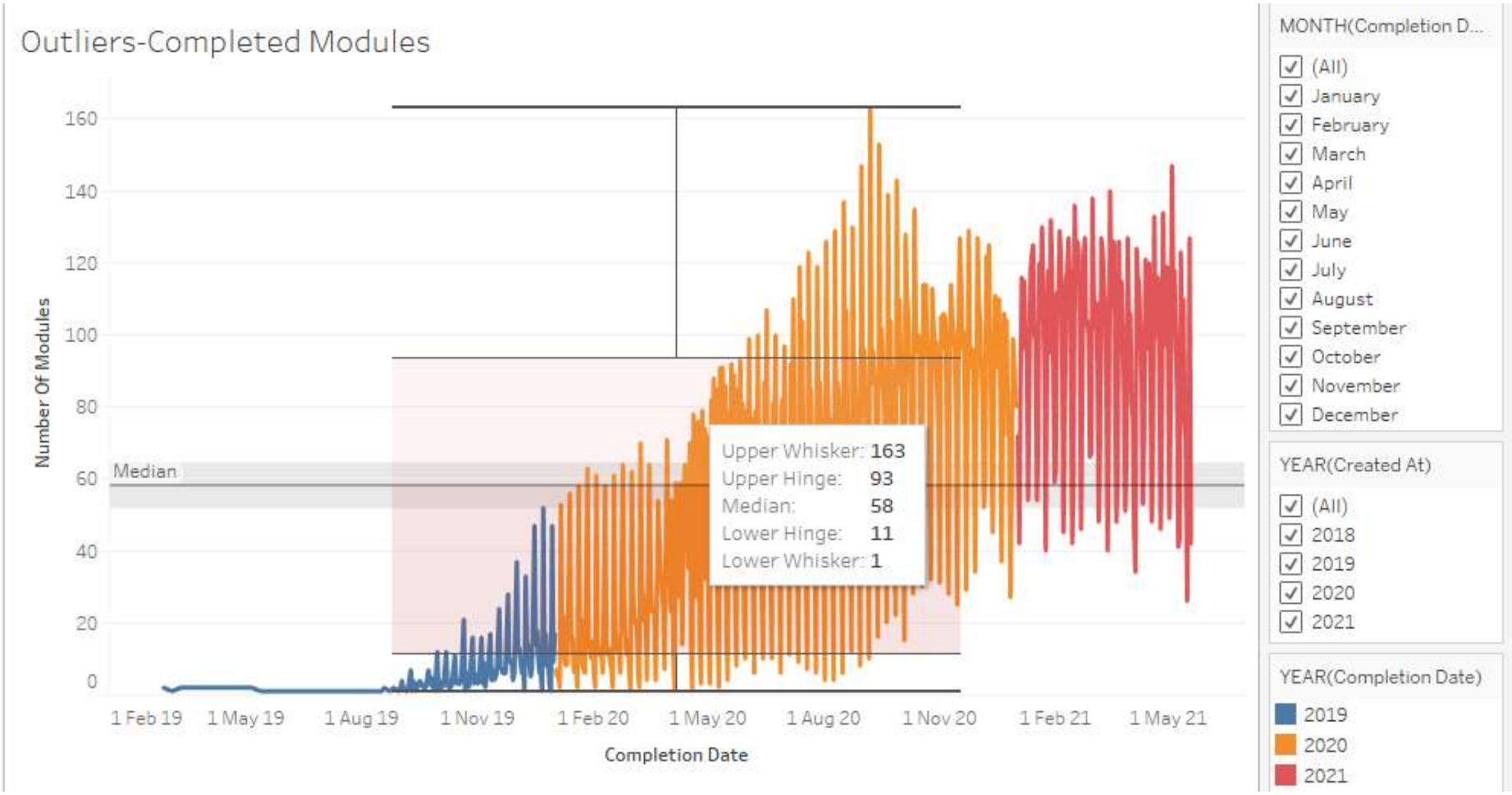
There are not null values, at least in the new database

| patient_id | clinic_id | clinic_title | CLINIC_GROUP | created_at | COMPLETION_DATE | NUMBER_OF_MODULES |
|---|---|---|---|---|---|---|

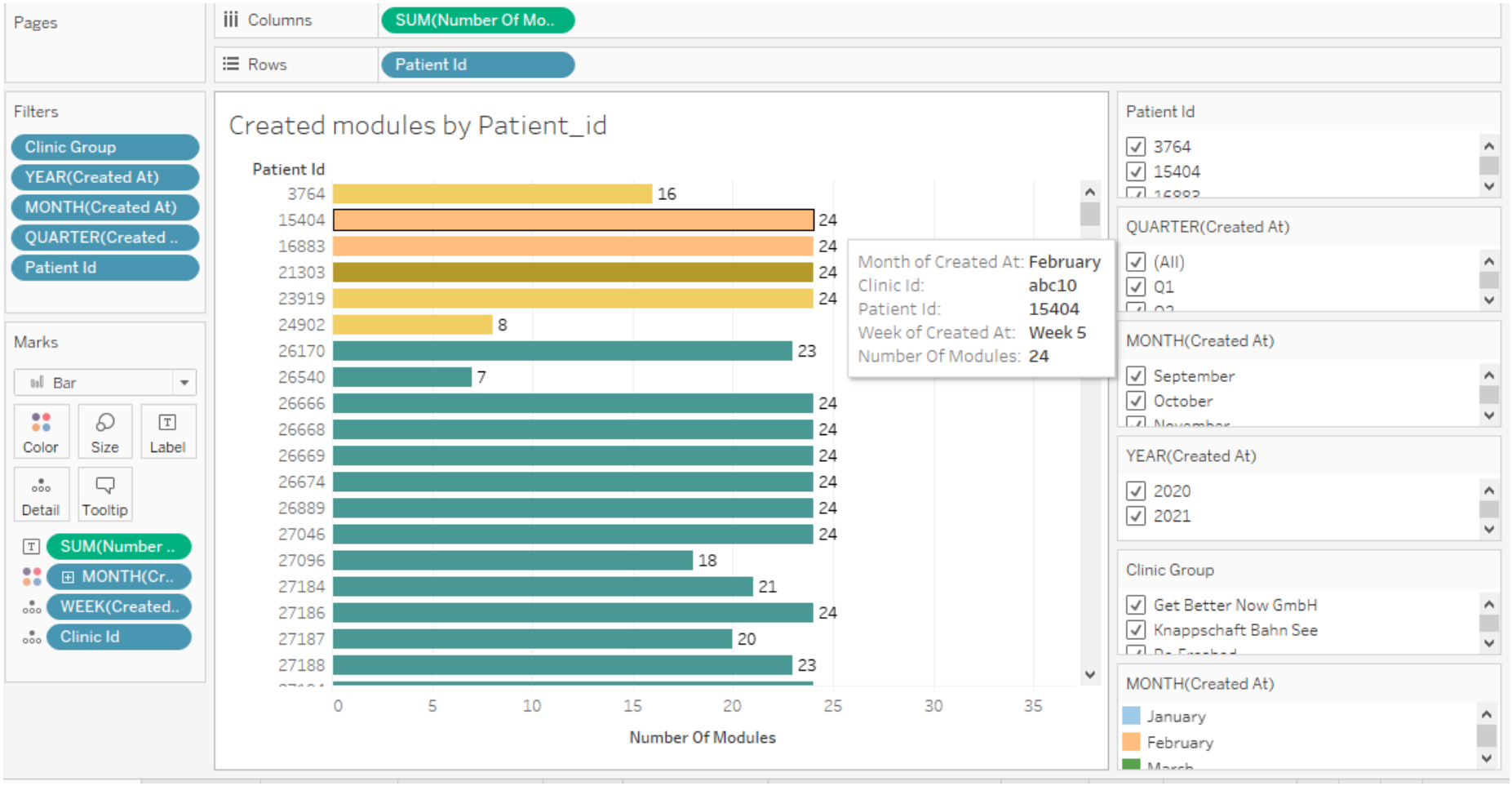
Result 42: 50 11:28:14 SELECT (select count(modules.patient_id) FROM assignment_data_modules ... 1 row(s) returned 0.094 sec / 0.000 sec

# Tableau

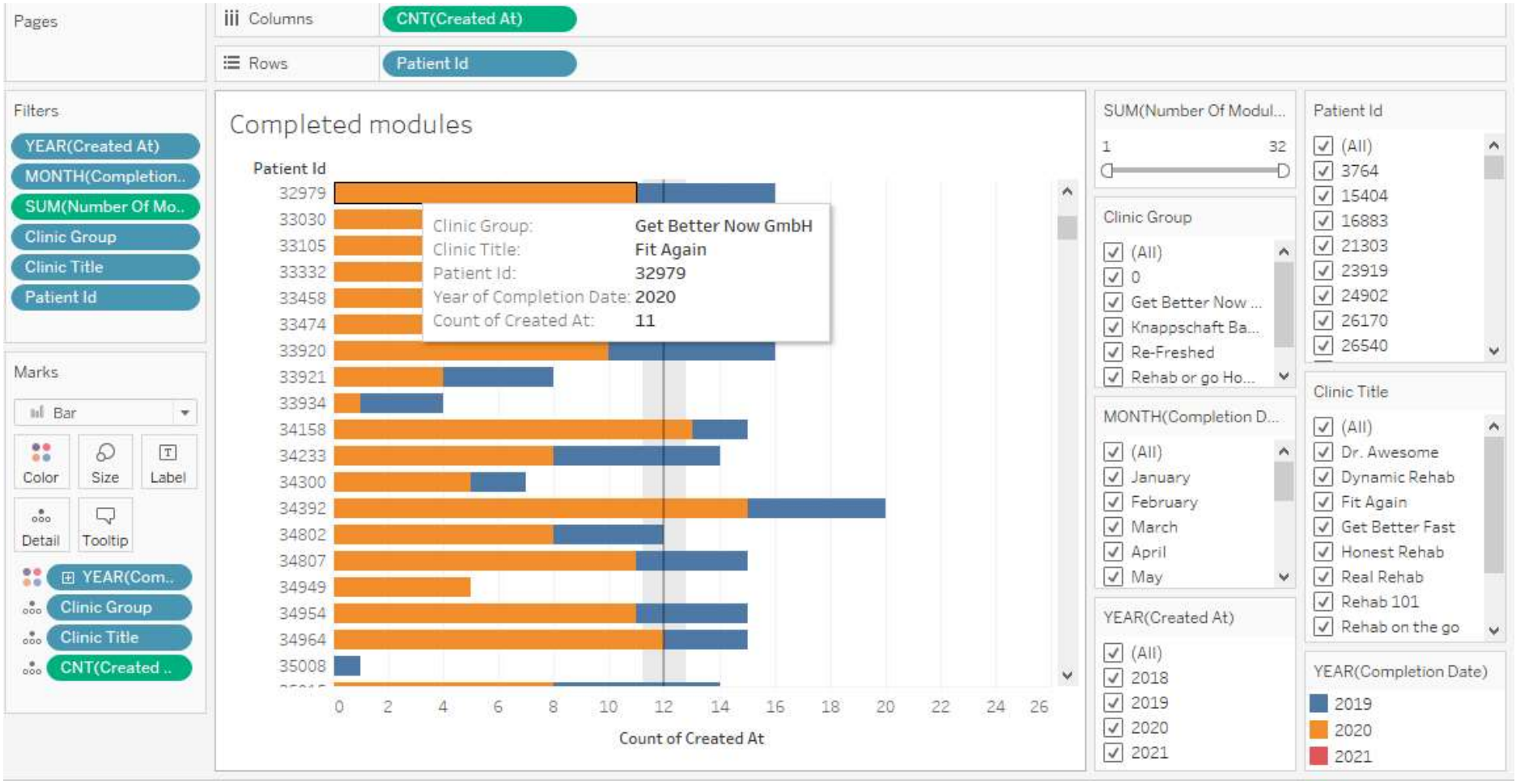
# Outliers: number of completed modules



# Visualizations: Created Modules

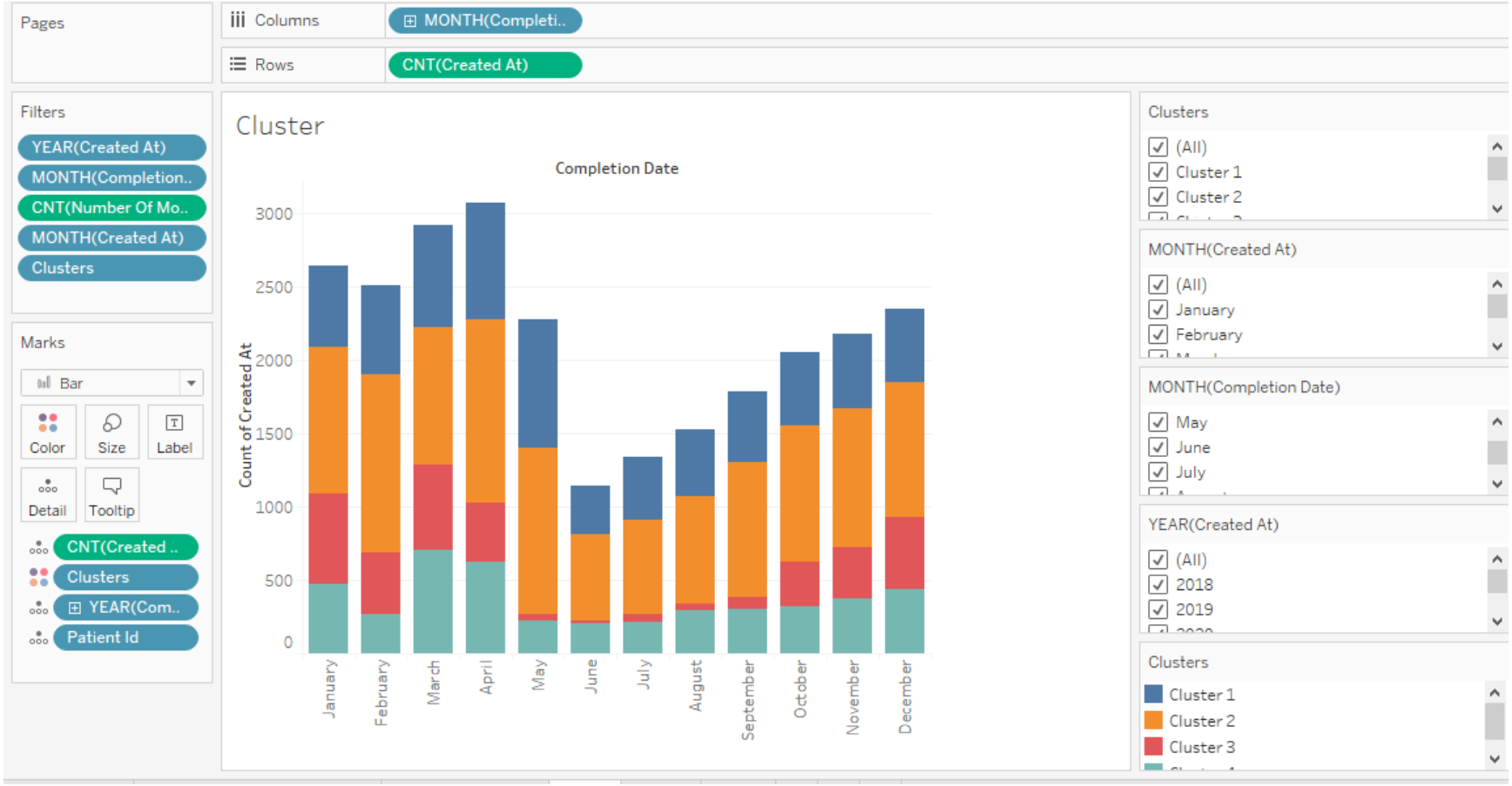


# Visualizations:
# Completed Modules vs created modules




Median value with 95% CI=12

# Clustering(1)



# Clustering(2)

Describe Clusters

Summary | Models

**Inputs for Clustering**

Variables: Count of Created At
Level of Detail: Month of Completion Date, Patient Id, Year of Completion Date
Scaling: Normalized

**Summary Diagnostics**

Number of Clusters: 4
Number of Points: 9253
Between-group Sum of Squares: 280.67
Within-group Sum of Squares: 28.997
Total Sum of Squares: 309.67

| Clusters | Number of Items | Centers Count of Created At |
|---|---|---|
| Cluster 1 | 4664 | 0.047146 |
| Cluster 2 | 3180 | 0.2812 |
| Cluster 3 | 515 | 0.63474 |
| Cluster 4 | 894 | 0.44444 |
| Not Clustered | 0 | |

Show scaled centers

Copy to Clipboard | Learn more about the cluster summary statistics | Close

Describe Clusters

Summary | Models

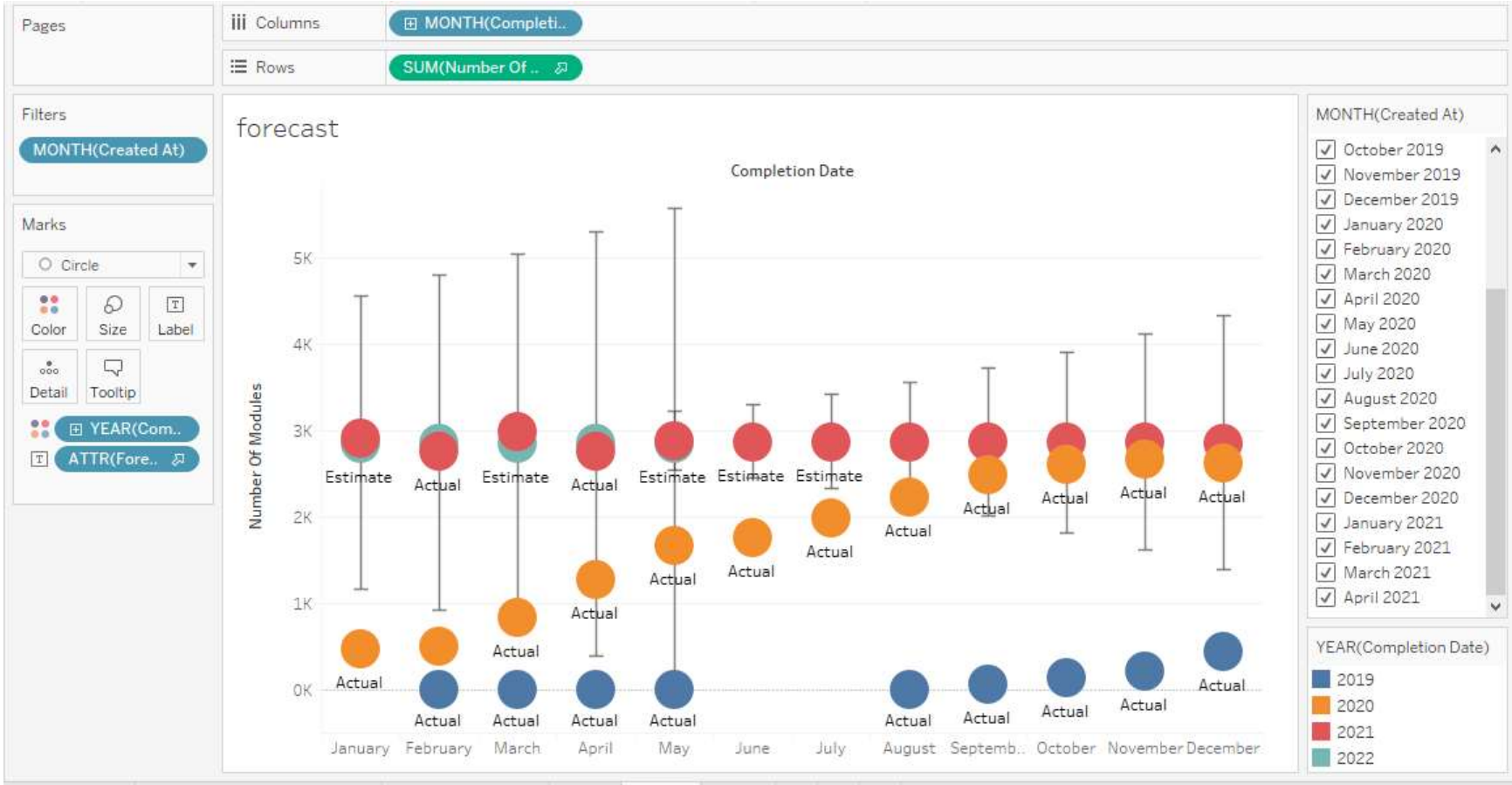
**Analysis of Variance:**

| Variable | F-statistic | p-value | Model Sum of Squares | Model DF | Error Sum of Squares | Error DF |
|---|---|---|---|---|---|---|
| Count of Created At | 2794.0 | 0.0 | 280.7 | 3 | 309.7 | 9249 |

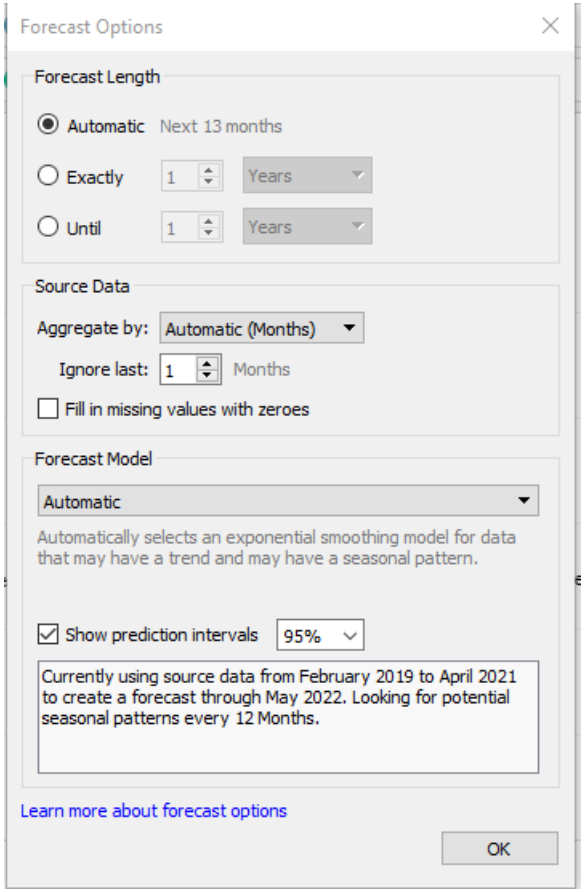
The p-value is inferior to 0.05, it means there is statistic difference between the clusters .

Copy to Clipboard | Learn more about the cluster model statistics | Close

# Forecasting (1)



# Forecasting (2)





Trend and Season depend very strong on the number of months (or other time of unit) in order to predict time series (how many modules will be done in 12 months)

There isn't a constant seasonality across the decomposition. The magnitude of the seasonal pattern in the data depends on the magnitude of the data, so therefore It's needed to have more data over time in order to calculate the seasonality

Describe Forecast

Summary | Models

All forecasts were computed using exponential smoothing.

**Sum of Number Of Modules**

| Model | | | Quality Metrics | | | | | Smoothing Coefficients | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Level | Trend | Season | RMSE | MAE | MASE | MAPE | AIC | Alpha | Beta | Gamma |
| Additive | Additive | None | 174 | 145 | 0.93 | 461.4% | 268 | 0.500 | 0.500 | 0.000 |

In Seasonality you can predict several years, for example, from 1950 to 1990. Meanwhile additive could be an example times series from one year to another

In the additive model, the behavior is linear where changes over time are consistently made by the same amount.

Copy to Clipboard | Learn more about the forecast models | Close