# Azrieli and Sons – Problem Log

In this doc, I'll record all data related issues I have encountered during the exercise.

## Taarif.csv

### Missing & Fishy Values

1. The table contains an empty row, probably a typo. **Deleted**
2. All values are missing for "bituch\_leumi", need to ask the customer for values. **See adjustment 10 and 11**
3. Electricity's extra mile price is missing, and weekend bonus seems like a mistake (70000%), need to ask the customer for values. **See adjustment 10**
4. Some values are 0 (weekend and night bonuses), it's a good idea to make sure there are no mistakes here. **See adjustment 10**

## Drivers\_with\_kviut.csv

### Indexing

1. Has an extra column with no name but with unique values, looks like an index. **Ignored**

### Missing & Fishy Values

1. Birthdate and gender have missing values (nan)
2. Gender has "none", "unknown" values. are they equivalent to nan or some 2019 gender-fluid progressive stuff? **See adjustment 1**
3. Birthdate has "1/1/1900" values. Usually means the date is unknown and was set to a default value. **See adjustment 2**

### Formatting

1. Birthdate: 01/01/1900, 01.01.1900, January 1, 1900. **Transformed to a single format**
2. Gender: 'F', 'f', 'girl', 'woman', 'female' (same for males). **Transformed to a single format**

## New\_drivers.csv

Note: This dataset wasn't mentioned in the instructions

### Indexing

1. Has an extra column with no name but with unique values, looks like an index. **Deleted**

### Missing & Fishy Values

1. Birthdate and gender have missing values (nan)
2. Gender has "none", "unknown" values. are they equivalent to nan or some 2019 gender-fluid progressive stuff? **See adjustment 1**
3. Birthdate has "1/1/1900" values. Usually means the date is unknown and was set to a default value. **See adjustment 2**

### Formatting

1. Birthdate: 01/01/1900, 01.01.1900, January 1, 1900 **Solved**
2. Gender: 'F', 'f', 'girl', 'woman', 'female' (same for males) **Solved**
3. Unlike Drivers\_with\_kviut.csv, it seems like vetek is in days (make sure it's not weeks/months) instead of years. **See adjustment 3**

## Trips

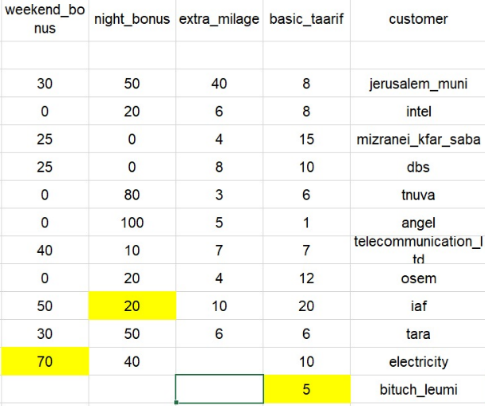
### Missing & Fishy Values

1. Some customers do not appear in Taarif.csv ("yes", "hot", "aminach") **See adjustment 8**
2. Missing values in start or end dates **See adjustment 14**
3. Missing values in BOTH start and end dates
4. Large values in ‘km’ (~10,000). Sometimes create odd average speed (such as 236 km/h).

### Indexing

1. Some files appear twice (same truck and month), content is also identical.

# Questions for Client and Adjustments Made Accordingly:

1. I assumed that "none", "unknown"… are equivalent to missing data and saved it as 'U'.
2. I assumed "1/1/1900" is equivalent to missing data and saved it as NaN.
3. Vetek in new drivers' table is in days.
4. Weekend is Friday (16:00) to Saturday (20:00).
5. What counts as "night"? 22:00 – 6:00
6. Which taarif is considered when more than one is relevant (weekend, extra, night)? For example: basic=20, night = 10% and weekend=40%, calculate as follows: 20\*1.4\*1.1 per km.
7. How to calculate mileage and wage when end or start time are missing? In order to complete missing dates, I used the average driving speed of a driver in a specific day (using known date and only if can be calculated. Else – overall average speed) and a known date to calculate the missing one.
8. Customers:
   1. Hot = telecommunication\_ltd
   2. Yes = dbs
   3. Aminach = mizranei\_kfar\_saba
9. A Trip can be partly night/weekend
10. Fixed taarif table, given by client:
11. In order to complete missing values in Taarif, I will use the mean value of basic\_taarif/extra\_milage and the known basic\_taarif. Same for unknown bonuses.
12. Is it OK to assume a driver will drive at the same speed at night and at day? Yes
13. Are all trips occurring during a single day (no cross-days trips)? No
14. How to handle trips that have no start & end date (suggestion: use the month that is in the filename and assume no bonuses of night/weekend for all trip)? As suggested.
15. How to handle trips that are longer than a day (in terms of rest, average speed…)?
16. Using this information:

* Total number of trips: 318181
* Total number of trips longer than 1000km: 4549 (1.4297%)
* Total number of trips longer than 3000km: 1105 (0.3473%)
* Total number of trips longer than 5000km: 40 (0.0126%)
* Total number of trips with average speed higher than 100: 3177 (0.9985%)
* Total number of trips with average speed higher than 150: 3119 (0.9803%)
* Total number of trips with average speed higher than 200: 2088 (0.6562%)
* Total number of trips with average speed higher than 225: 781 (0.2455%)
* Total number of trips longer than 8 hours: 3438 (1.0805%)
* Total number of trips longer than 12 hours: 3248 (1.0208%)
* Total number of trips longer than 24 hours: 3223 (1.0129%)
* Total number of trips longer than 48 hours: 1 (0.0003%)
* Dropping all trips using the lowest thresholds will remove 2.0667% of the data

Can I drop trips using the lowest thresholds (1000km, 100km/h, and 8 hours)?

1. Trips with missing date and fishy km will be removed (because we can’t evaluate their dates using the distance and can’t evaluate the distance using the dates.

# Meetings with Client

**26.3.2020:**

* Answered question 7 regarding the way of handling a single missing date.
* Answered question 14 regarding handling trips with missing dates (both).
* Discussed question 15. Need to have precise questions, present relevant graphs (use plotly!) and come with concrete suggestions.
  + After talking to the client, decided that each trip that is over 1000 km is fishy and that its km should be updated using the average speed table.
  + Can drop trips that are longer than 8 hours
  + Can drop trips that are at average speed of more than 100 km/h