Data Issues

Taarif:

1. Bituch Leumi and Electricity have missing values and values which seem too big.

**Solution**: Viktor told me those were old values and to put 5 for both basic and extra rate and no bonuses.

1. The night bonus of IAF and Angel and the extra milage rate of Jerusalem Muni seem too high.

**Solution**: They have an extra 0, divide them by 10.

1. The companies Hot, Aminach, and Yes from the trips tables don’t show up in the rate table.

**Solution**: some customers have two names, as shown in the below chart, I changed the names in the Taarif tables to the other set of names.

|  |  |
| --- | --- |
| Taarif Table Name | Trips Customer Name |
| mizranei\_kfar\_saba | aminach |
| dbs | yes |
| telecommunication\_ltd | hot |

Drivers:

1. Young drivers have much higher Vetek values then senior drivers.

**Solution**: the units for young drivers are in days and the units for seniors are years, I divided the values for young drivers by 365.

1. Some of the drivers don’t have any gender.

**Solution**: filled in ‘unknown’.

1. It is not always possible to compute a driver’s age, some drivers don’t have a birthdate, others have 1.1.1900 as a placeholder.

**Solution**: I set the age of all drivers without a valid date to nan.

1. Dates come in several formats: 1.1.2022, 1/1/2022 and January 1, 2022. **Solution**: Pandas to\_date fixes this in consistency.

Trips:

1. There is more then one file for some trucks and some months.

**Solution**: the files are always duplicates, only one needs to be handled.

1. There are many entries which don’t have end time or start time, some don’t have both.

**Solution**: when both are not present, the start time is set to be the first day of the month at 12 pm. When one is set, we find the trip’s duration by dividing the trip’s distance by the driver’s average speed and using that to find the missing start/end time.

1. Bonus is per kilometer, but how can we tell how many kilometers were driven during night/weekend?

**Solution**: assume distance travelled is uniform, the time travelled at night is the total trip’s distance, times the fraction of the time driven at night.

1. Some trips have speeds which are too slow and some trips have speeds which are too high or are too long.

**Solution**: For trips with speeds which are too low, the driver’s average speed was used to find an approximated drive duration and end time. For trips which are too long or too quick, the km was replaced with the driver’s average drive distance, and the drivers average speed was used.