Bike Sharing - Demand Analysis & Prediction

Pedalmetrics Consulting

Miriam Godinez
Flora Kwong

Overview

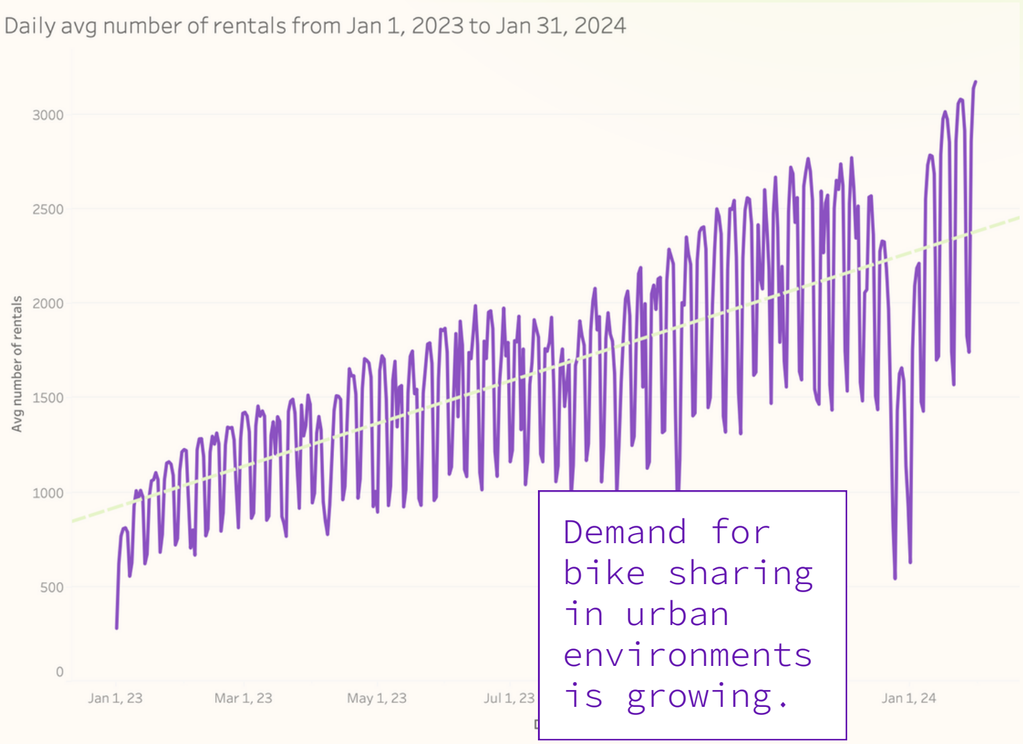
1. Introduction
2. Approach
3. User and market analysis
4. Data modelling/Machine Learning
5. Forecasting demand
6. Learnings & challenges
7. Summary

01 - Introduction

Case study:
Ecobici in Mexico City

Residents of CDMX 158 hours per year commuting (Ortego et al 2021).
Transportation makes up ~14 % of global GHG emissions (Statista, 2022)

PedalMetrics can help bike sharing companies prepare:
- **User and market analysis**
- **Forecasting inventory needs**



02 - Approach

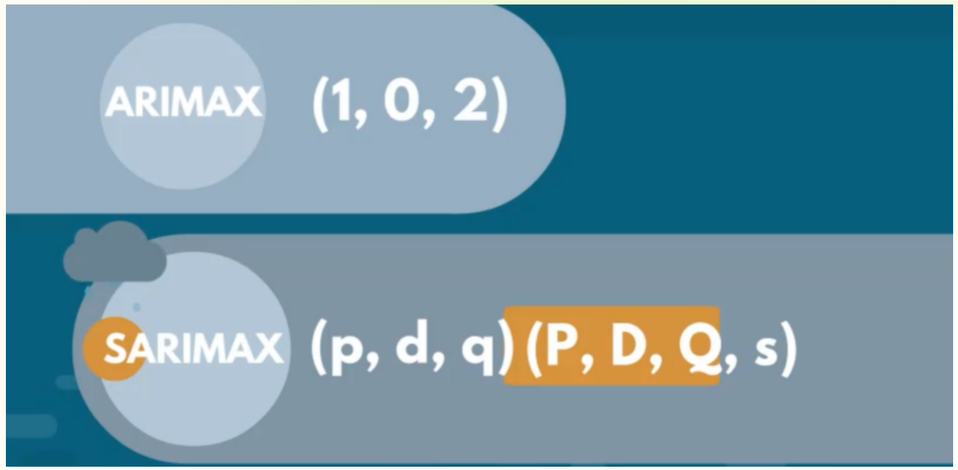
1. Wrangle & preprocess

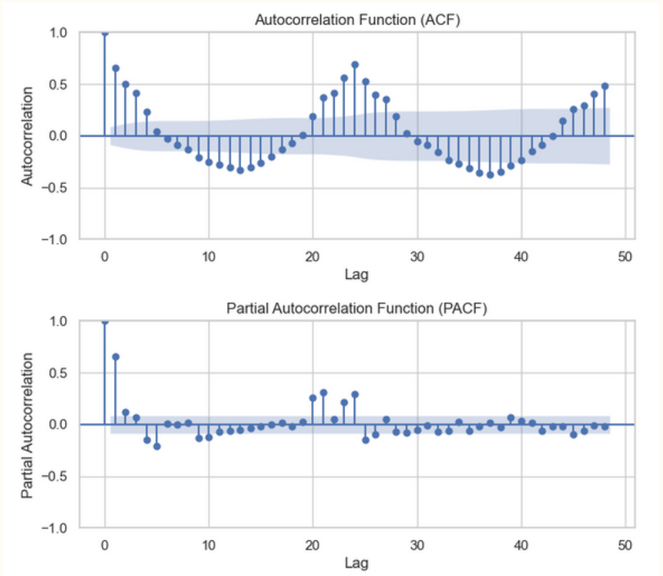
- Historical bike sharing data cleaned, transformed and aggregated by hour.
- Merge with historical hourly weather information.
- Split the data by bike pick-up station.

2. Exploratory data analysis

- Determine general patterns and trends in your users and demand.
- Explore potential features for our data modelling.

02 - Approach





3. Machine learning & modelling (Time series analysis-hourly)

- Prepare the data for SARIMAX time-series modelling with exogenous variables.
- Tune our SARIMAX hyperparametres through a grid search.
- Fit SARIMAX model on our train and test data.
- Scale for all pick-up stations and continuously update with new data.

4. Predict demand

- Extract weather forecast data for desired forecast period
- Forecast the number of bikes needed at each rental station on an hourly basis, for a semi-monthly or monthly time period.

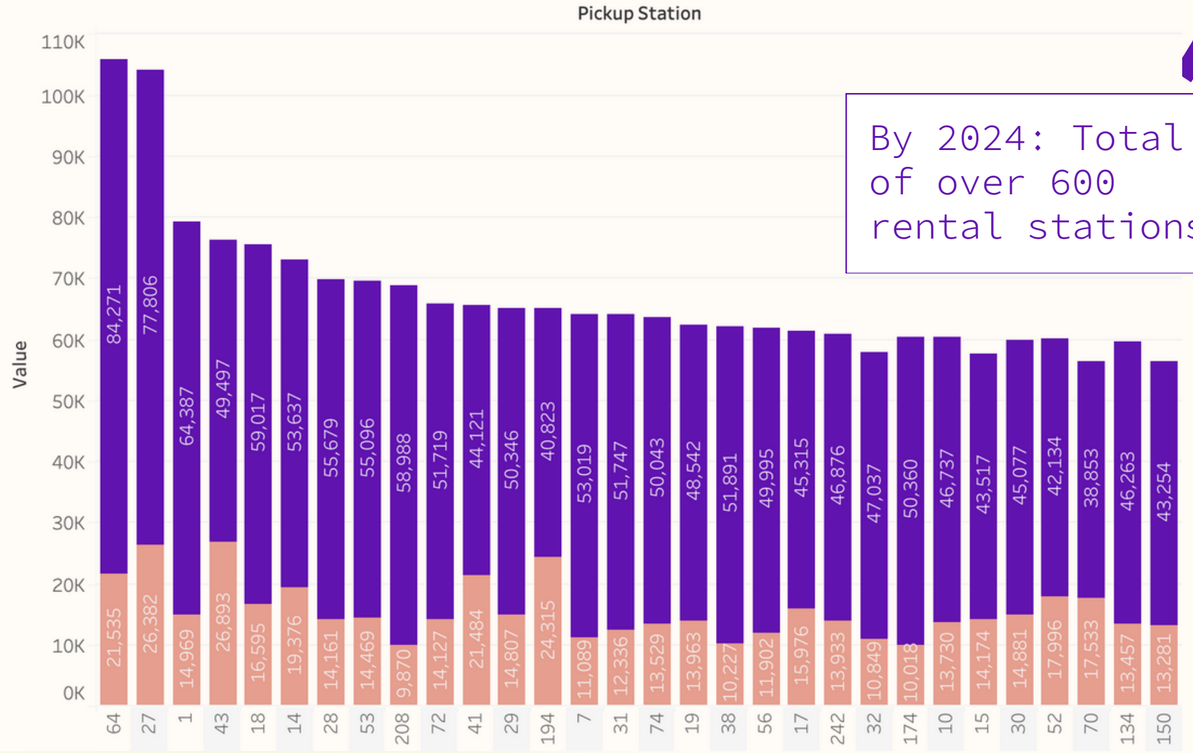
03 - Ecobici: user and market analysis

Since Feb 2010:

- Cumulative rides: ~95 Million
- Cumulative Users: ~1 Million (Ecobici, 2024)

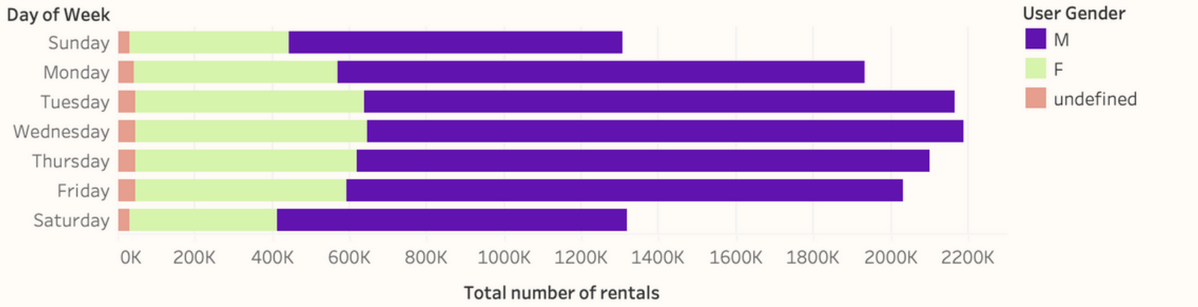
Our analysis focus:
01/01/2023 to 31/01/2024

Total number of rentals for top 30 pick-up stations (business vs. non-business days) from Jan 2023 to Jan 2024



By 2024: Total of over 600 rental stations

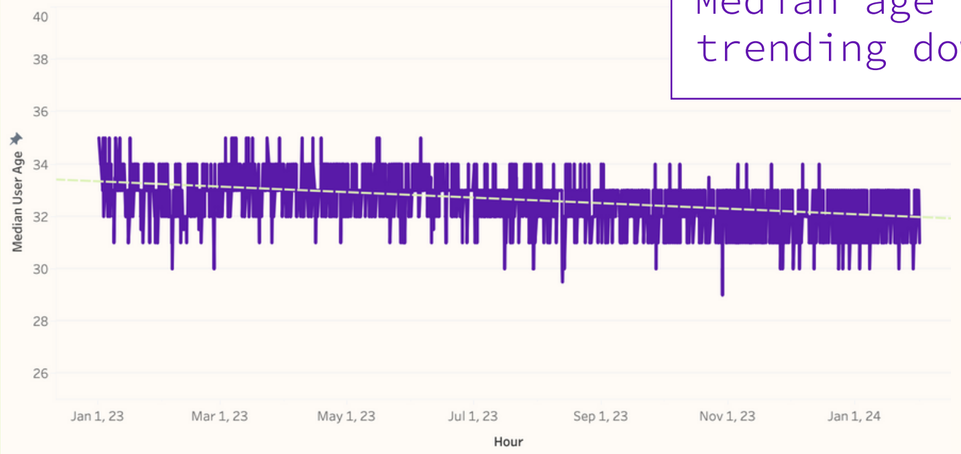
Total number of rentals per day of week by gender



Majority male customers, most use on Tue, Wed, Thu

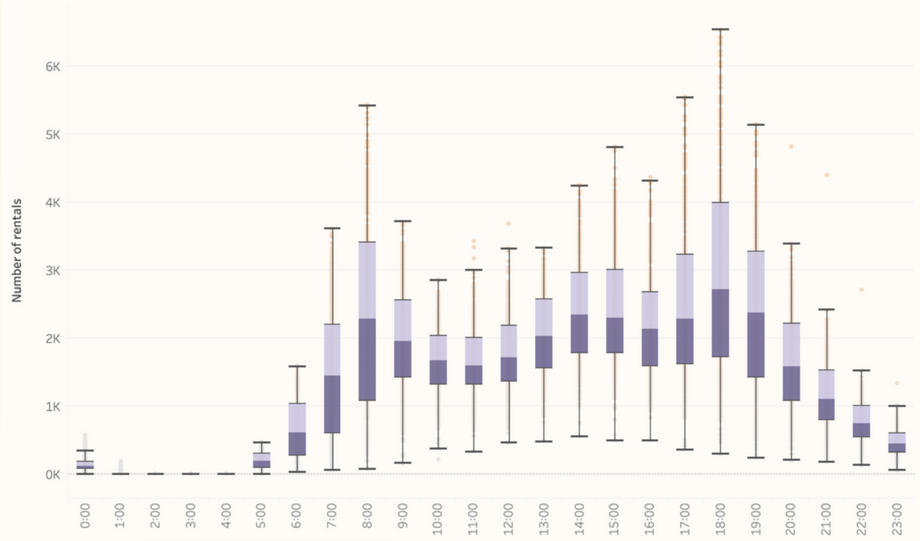
03 - Ecobici: user and market analysis

Median user age trend



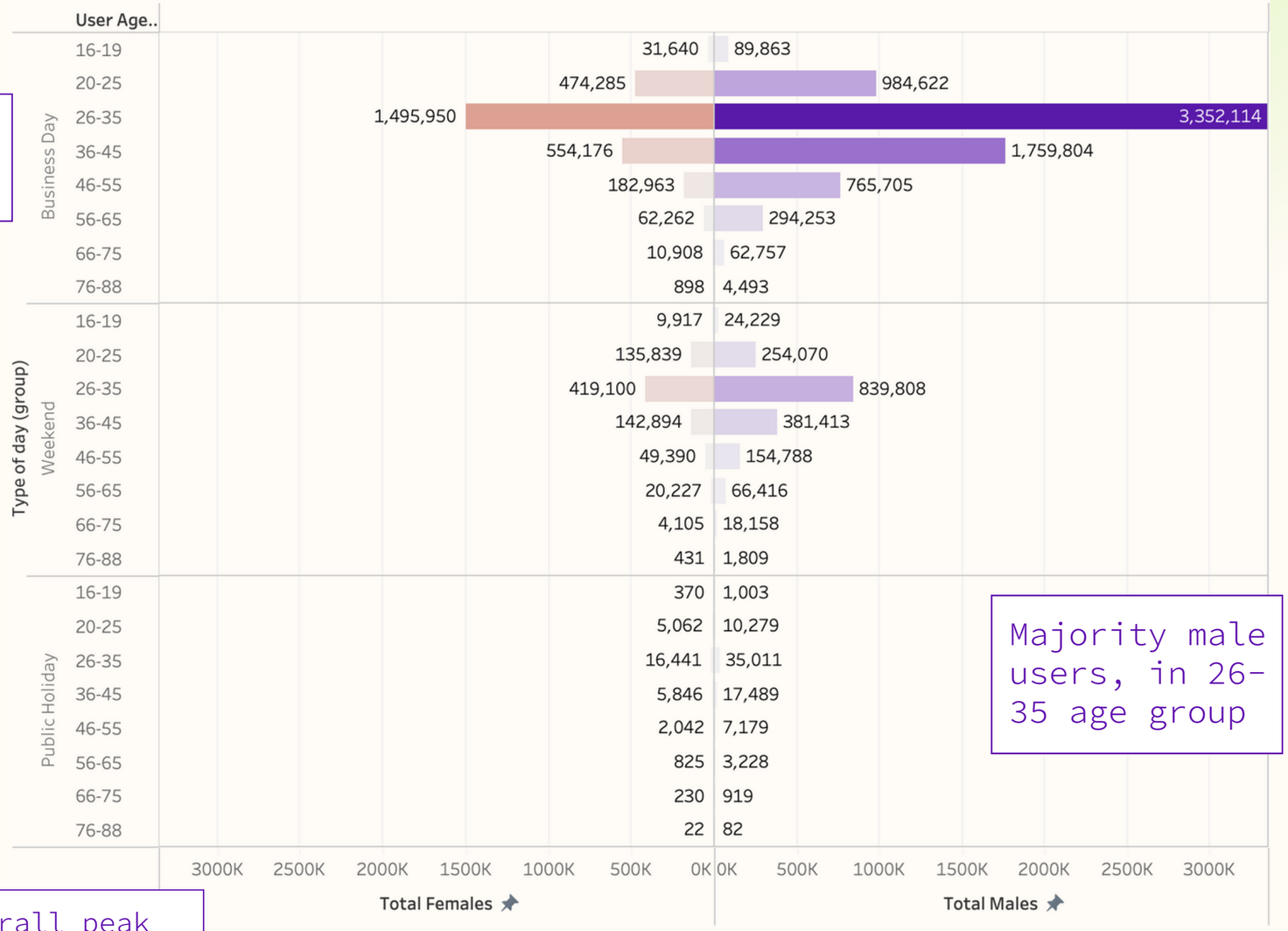
Median age trending down

Distribution of avg number of rentals per hour



Overall peak demand on weekdays ~8 am and ~6 pm

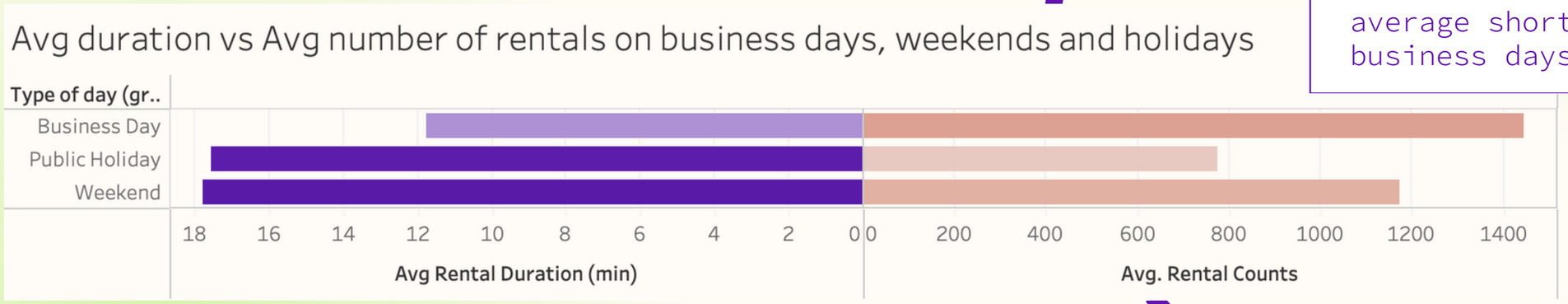
Number of rentals per type of day by age group (filtered ages 16-88) and gender

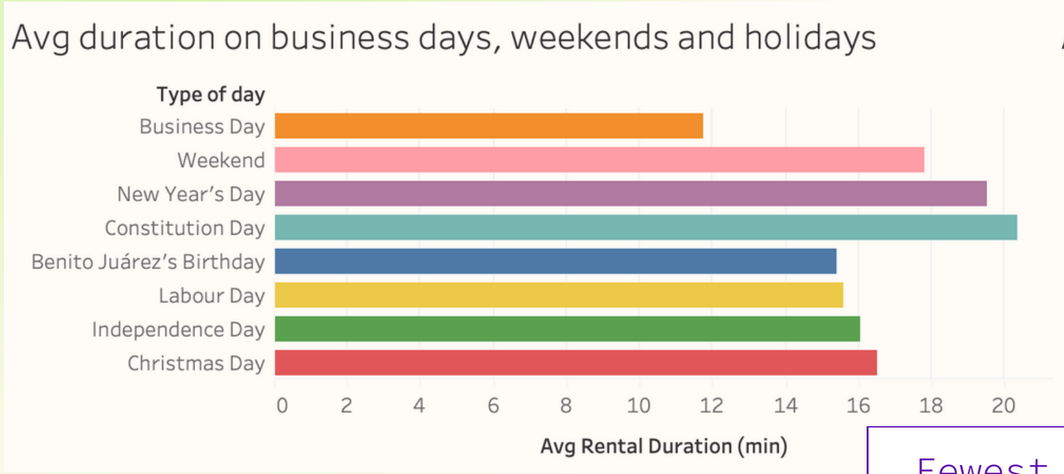


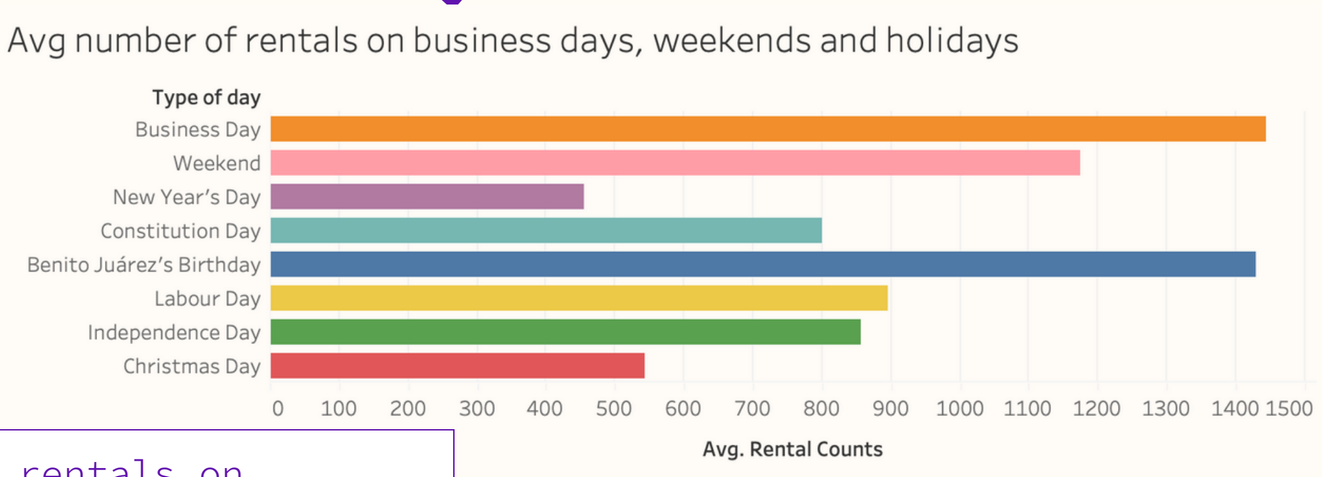
Majority male users, in 26-35 age group

03 - Ecobici: User and market analysis

More rentals on business days than weekends and holidays. In contrast, rental durations are on average shorter on business days







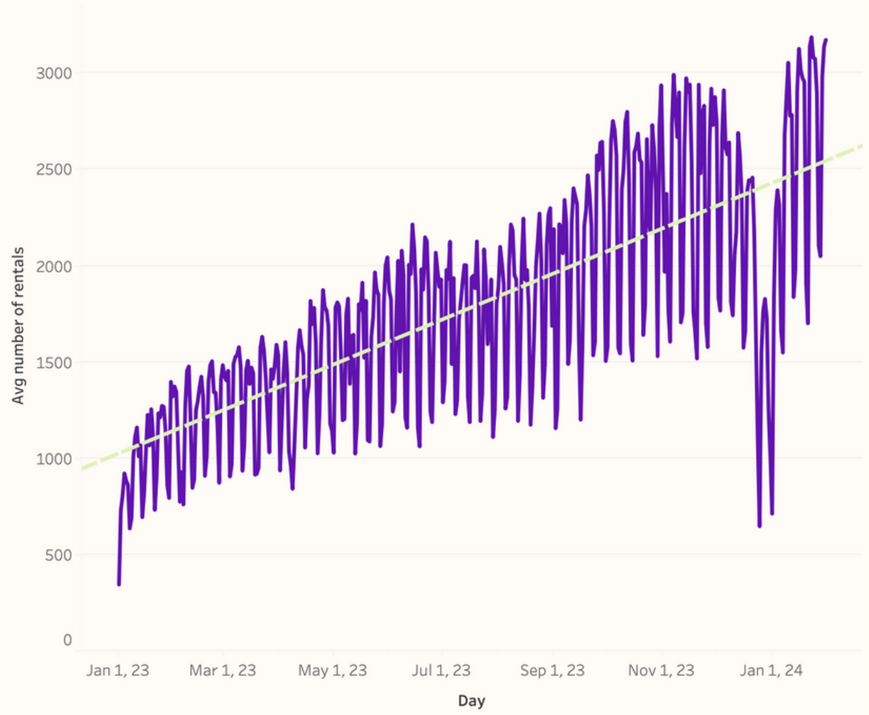
Fewest rentals on Christmas and New Year's

04 - Data Modelling & Machine Learning

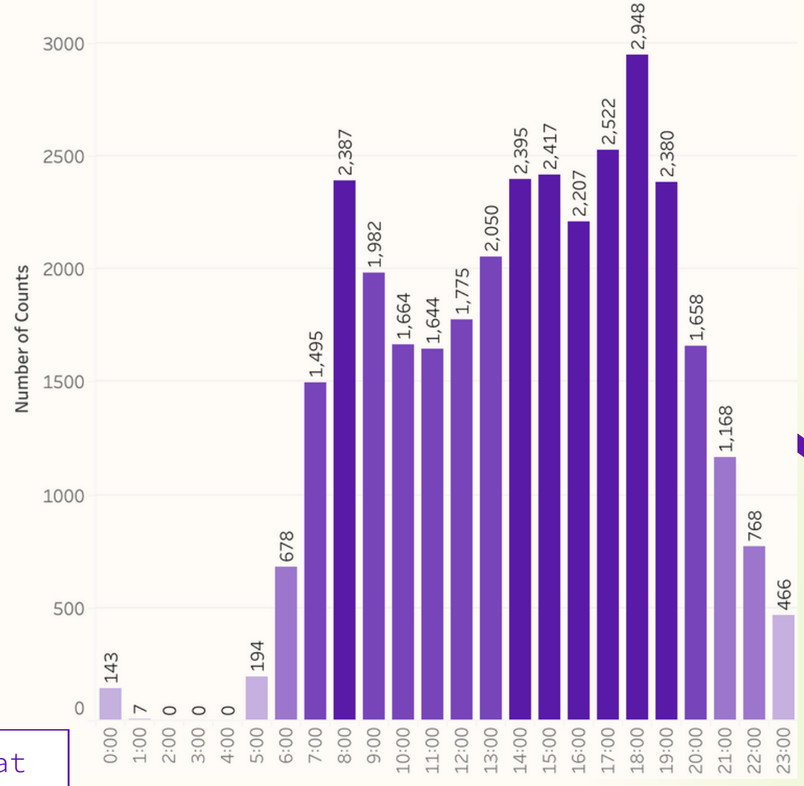
Zooming in on rental station #6

44th in top total rental counts

Daily avg number of rentals for **station 6** from Jan 1, 2023 to Jan 31, 2024



Avg number of rentals per hour for **station #6**



Overall growing demand at station 6, with peak hours of demand on a workday 8am, 2-3pm and 6pm

Choosing a model:

SARIMAX
Time-series data
Seasonality (24 hours)
Exogenous variables (weather, is_business_day)
Can be applied individually to each of our 644 rental stations

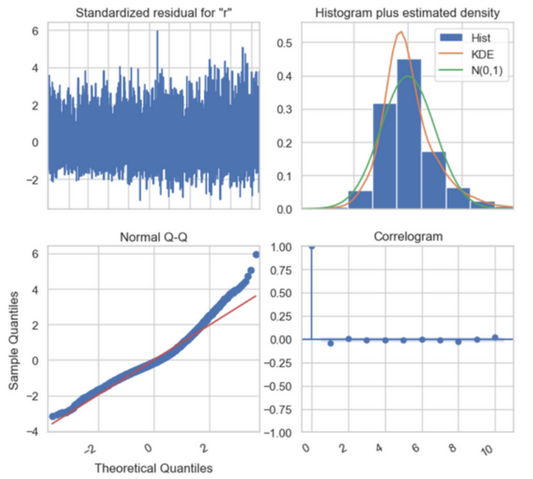
Tuning the model:

Grid search: automated hyperparametre tuning
Scaling/transforming our exogenous variables
Feature selection: checking coefficients and associated p-value

Model 1:
(1, 0, 1)(1, 0, 1, 24)
All exogenous variables
No scaling/transforming

	coef	std err	z	P>|z|	[0.025	0.975]
is_business_day	2.4904	0.148	16.833	0.000	2.200	2.780
temperature_C	-0.1322	0.074	-1.786	0.074	-0.277	0.013
rel_humidity_perc	0.0522	0.003	16.920	0.000	0.046	0.058
feels_like_temp_C	0.1407	0.071	1.975	0.048	0.001	0.280
rain_mm	-0.9166	0.057	-16.011	0.000	-1.029	-0.804
cloud_cover_perc	-0.0127	0.002	-6.382	0.000	-0.017	-0.009
wind_speed_kmh	-0.1454	0.017	-8.744	0.000	-0.178	-0.113
is_day	3.5870	0.268	13.360	0.000	3.061	4.113
ar.L1	0.5726	0.029	19.936	0.000	0.516	0.629
ma.L1	-0.2741	0.032	-8.647	0.000	-0.336	-0.212
ar.S.L24	0.8006	0.016	51.150	0.000	0.770	0.831
ma.S.L24	-0.6011	0.022	-27.463	0.000	-0.644	-0.558
sigma2	11.1444	0.167	66.674	0.000	10.817	11.472

Ljung-Box (L1) (Q):	13.01	Jarque-Bera (JB):	1799.93
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.55	Skew:	0.85
Prob(H) (two-sided):	0.00	Kurtosis:	4.67

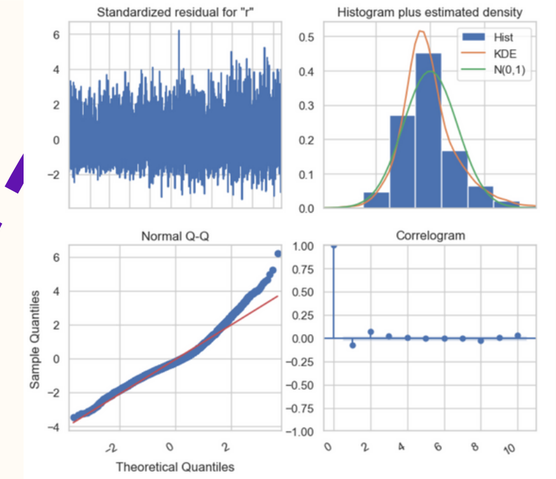


mean absolute error: 3.46

Model 2:
(1, 0, 0)(1, 0, 1, 24)
Remove temperature feature
No scaling/transforming

	coef	std err	z	P>|z|	[0.025	0.975]
is_business_day	2.2558	0.109	20.763	0.000	2.043	2.469
rel_humidity_perc	0.0560	0.003	20.167	0.000	0.051	0.061
rain_mm	-1.0074	0.049	-20.549	0.000	-1.103	-0.911
cloud_cover_perc	-0.0136	0.002	-7.859	0.000	-0.017	-0.010
wind_speed_kmh	-0.0809	0.010	-7.961	0.000	-0.101	-0.061
is_day	4.0211	0.206	19.547	0.000	3.618	4.424
ar.L1	0.2924	0.010	30.545	0.000	0.274	0.311
ar.S.L24	0.8730	0.009	95.933	0.000	0.855	0.891
ma.S.L24	-0.6467	0.015	-43.841	0.000	-0.676	-0.618
sigma2	10.0309	0.135	74.562	0.000	9.767	10.295

Ljung-Box (L1) (Q):	39.26	Jarque-Bera (JB):	1865.73
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.56	Skew:	0.83
Prob(H) (two-sided):	0.00	Kurtosis:	4.77

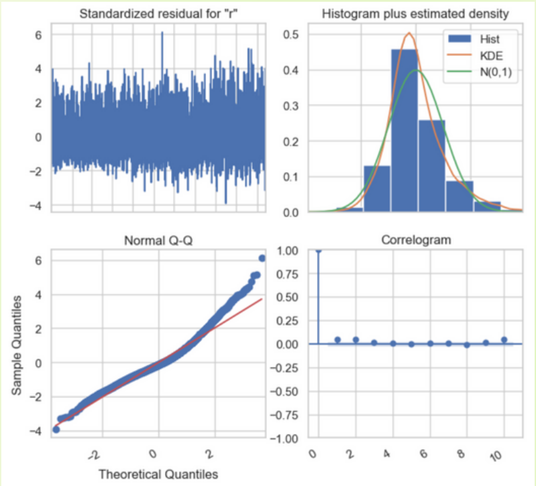


mean absolute error: 3.5

Model 5:
(1, 0, 1)(1, 0, 1, 24)
Remove cloud cover feature
Scaled exog with MinMax

	coef	std err	z	P>|z|	[0.025	0.975]
is_business_day	2.1434	0.113	18.929	0.000	1.921	2.365
temperature_C	-9.8342	2.781	-3.536	0.000	-15.285	-4.383
rel_humidity_perc	4.4422	0.346	12.833	0.000	3.764	5.121
feels_like_temp_C	8.9463	2.535	3.529	0.000	3.978	13.914
rain_mm	-10.6814	0.526	-20.300	0.000	-11.713	-9.650
wind_speed_kmh	-4.7220	0.642	-7.360	0.000	-5.980	-3.464
is_day	4.3208	0.272	15.860	0.000	3.787	4.855
ar.L1	0.5126	0.041	12.647	0.000	0.433	0.592
ma.L1	-0.3200	0.044	-7.275	0.000	-0.406	-0.234
ar.S.L24	0.8460	0.011	80.181	0.000	0.825	0.867
ma.S.L24	-0.6134	0.016	-38.029	0.000	-0.645	-0.582
sigma2	10.3652	0.144	71.877	0.000	10.083	10.648

Ljung-Box (L1) (Q):	14.68	Jarque-Bera (JB):	1727.80
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.56	Skew:	0.81
Prob(H) (two-sided):	0.00	Kurtosis:	4.67



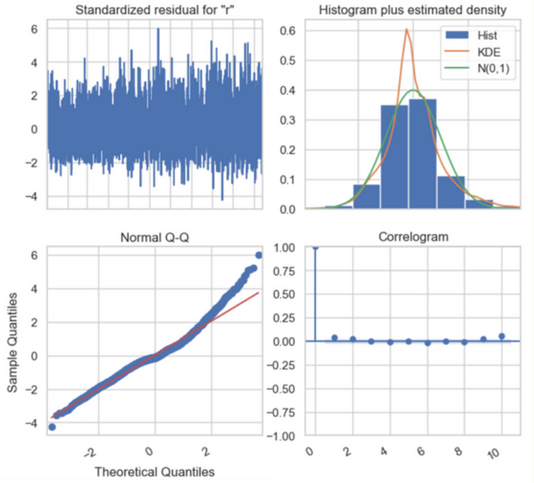
mean absolute error: 3.34

04 – Data Modelling & Machine Learning

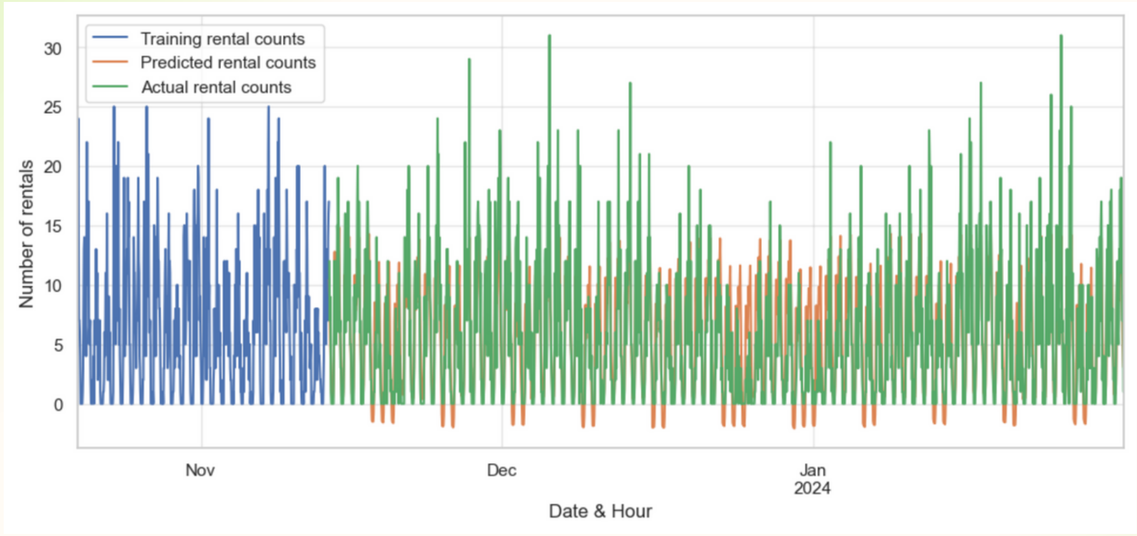
Model 999999:
(2, 0, 1)(0, 1, 2, 24)
Remove cloud cover feature
Scaled exog with MinMax

	coef	std err	z	P>|z|	[0.025	0.975]
is_business_day	2.4036	0.090	26.774	0.000	2.228	2.580
temperature_C	3.9141	3.556	1.101	0.271	-3.056	10.884
rel_humidity_perc	0.0695	0.713	0.098	0.922	-1.327	1.466
feels_like_temp_C	-1.9385	2.949	-0.657	0.511	-7.718	3.841
rain_mm	-0.1876	0.988	-0.190	0.849	-2.123	1.748
wind_speed_kmh	-0.1315	0.702	-0.187	0.851	-1.508	1.245
is_day	-0.9078	0.713	-1.273	0.203	-2.306	0.490
ar.L1	0.2661	0.205	1.298	0.194	-0.136	0.668
ar.L2	0.0301	0.025	1.227	0.220	-0.018	0.078
ma.L1	-0.1700	0.205	-0.829	0.407	-0.572	0.232
ma.S.L24	-0.7376	0.009	-83.352	0.000	-0.755	-0.720
ma.S.L48	-0.1728	0.009	-19.162	0.000	-0.190	-0.155
sigma2	8.5257	0.101	84.404	0.000	8.328	8.724

Ljung-Box (L1) (Q):	10.65	Jarque-Bera (JB):	1543.27
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.49	Skew:	0.60
Prob(H) (two-sided):	0.00	Kurtosis:	4.86



mean absolute error: 2.53

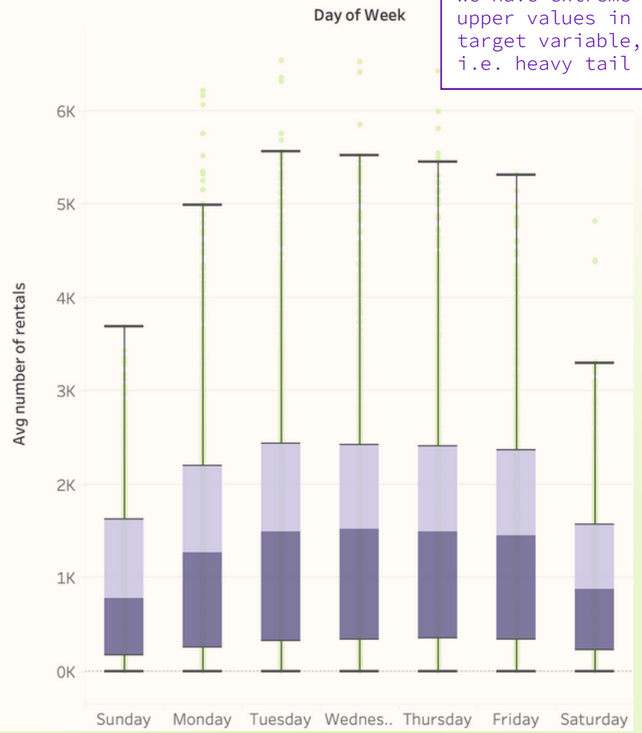


Assessing the (final) model:

- p-value of coefficients > 0.05: should not be rejecting H0 that the coefficients have an effect. BUT model performed worse when removing coefficients.
- Prob(Q): 0.00 < 0.05. Residuals are correlated, so further fine-tuning can be done
- Prob(JB):0.00 < 0.05. Residuals are random, i.e. not normally distributed.
- A taller KDE curve confirms that our target variable has more extreme values and is more heavily concentrated around the mean. Scaling the target variable didn't seem to help though.

Distribution of avg number of rentals per day of week

We have extreme upper values in our target variable, i.e. heavy tail



05 - Forecasting demand

Key factors that influence bike rental demand:

- Temperature
- Business days

```
In [44]: 1 retrieve_forecast(search_datetime).head(25)
```

Out[44]:

hour	temperature_C	rel_humidity_perc	feels_like_temp_C	rain_mm	wind_speed_kmh	is_day	is_business_day	rental_prediction
2024-02-01 00:00:00	21.3	21.0	18.4	0.0	3.3	1	1	1.0
2024-02-01 01:00:00	19.5	26.0	15.4	0.0	13.0	0	1	0.0
2024-02-01 02:00:00	18.6	29.0	16.2	0.0	2.5	0	1	0.0
2024-02-01 03:00:00	17.0	40.0	13.6	0.0	12.3	0	1	0.0
2024-02-01 04:00:00	14.8	53.0	12.9	0.0	5.0	0	1	0.0
2024-02-01 05:00:00	14.6	51.0	12.6	0.0	4.3	0	1	1.0
2024-02-01 06:00:00	13.0	56.0	10.9	0.0	5.2	0	1	5.0
2024-02-01 07:00:00	12.1	58.0	9.7	0.0	6.9	0	1	8.0
2024-02-01 08:00:00	12.2	56.0	10.3	0.0	2.2	0	1	9.0
2024-02-01 09:00:00	10.3	68.0	8.2	0.0	5.4	0	1	8.0
2024-02-01 10:00:00	9.3	70.0	7.2	0.0	5.1	0	1	6.0
2024-02-01 11:00:00	9.1	70.0	7.2	0.0	3.2	0	1	6.0
2024-02-01 12:00:00	9.2	66.0	7.2	0.0	2.5	0	1	6.0
2024-02-01 13:00:00	9.1	71.0	7.6	0.0	0.7	0	1	8.0
2024-02-01 14:00:00	9.1	71.0	7.5	0.0	1.4	1	1	11.0
2024-02-01 15:00:00	12.5	57.0	10.8	0.0	2.4	1	1	14.0
2024-02-01 16:00:00	15.6	44.0	13.7	0.0	2.5	1	1	9.0
2024-02-01 17:00:00	17.6	36.0	16.1	0.0	1.8	1	1	13.0
2024-02-01 18:00:00	19.1	31.0	17.6	0.0	2.6	1	1	15.0
2024-02-01 19:00:00	19.6	30.0	15.5	0.0	15.2	1	1	9.0
2024-02-01 20:00:00	18.1	37.0	13.7	0.0	19.1	1	1	4.0
2024-02-01 21:00:00	19.1	33.0	15.8	0.0	11.5	1	1	5.0

06 - Learnings & challenges

- Understanding time-series modelling, seasonality, and the use of exogenous variables
- For a complete analysis, each station requires its own model for pick-up and return trends.
- Realizing later on that we need to scale/transform our exogenous (and maybe target) variables...
  - and realizing that we need to save both the scaler and the model using pickle
- Grid search and modelling can easily be scaled with a for-loop for each of the 600+ rental stations.
- Computer capacity: retain our services and we can invest in more powerful computing on GCP or AWS
- Time constraints to create functions to import new data and automate wrangling + forecasting

07 - Conclusions

- The trend is bike-sharing Ecobici system is increasing, which indicates the need to expand to keep promoting sustainable transportation.
- ~70% of the rides are done by males, interesting; potential for further analysis of gender imbalance.
- Invest in us for scaling of the model, including:
  - Deeper analysis for marketing campaigns
  - Predictions for bike returns
  - Predictions for all bike stations
  - Predictions for other bike-sharing cities
- Use our forecasting model to ensure there are always enough bikes at each station!

Thanks

References:

Statista (2022): https://www.statista.com/topics/7476/transportation-emissions-worldwide/#topicOverview

Ortego et al (2021): https://link.springer.com/chapter/10.1007/978-3-030-69136-3_15

Ecobici (2024): https://ecobici.cdmx.gob.mx/en/statistics/

https://ecobici.cdmx.gob.mx/en/open-data/