DATA EXPLORATION

A Characteristics of the data

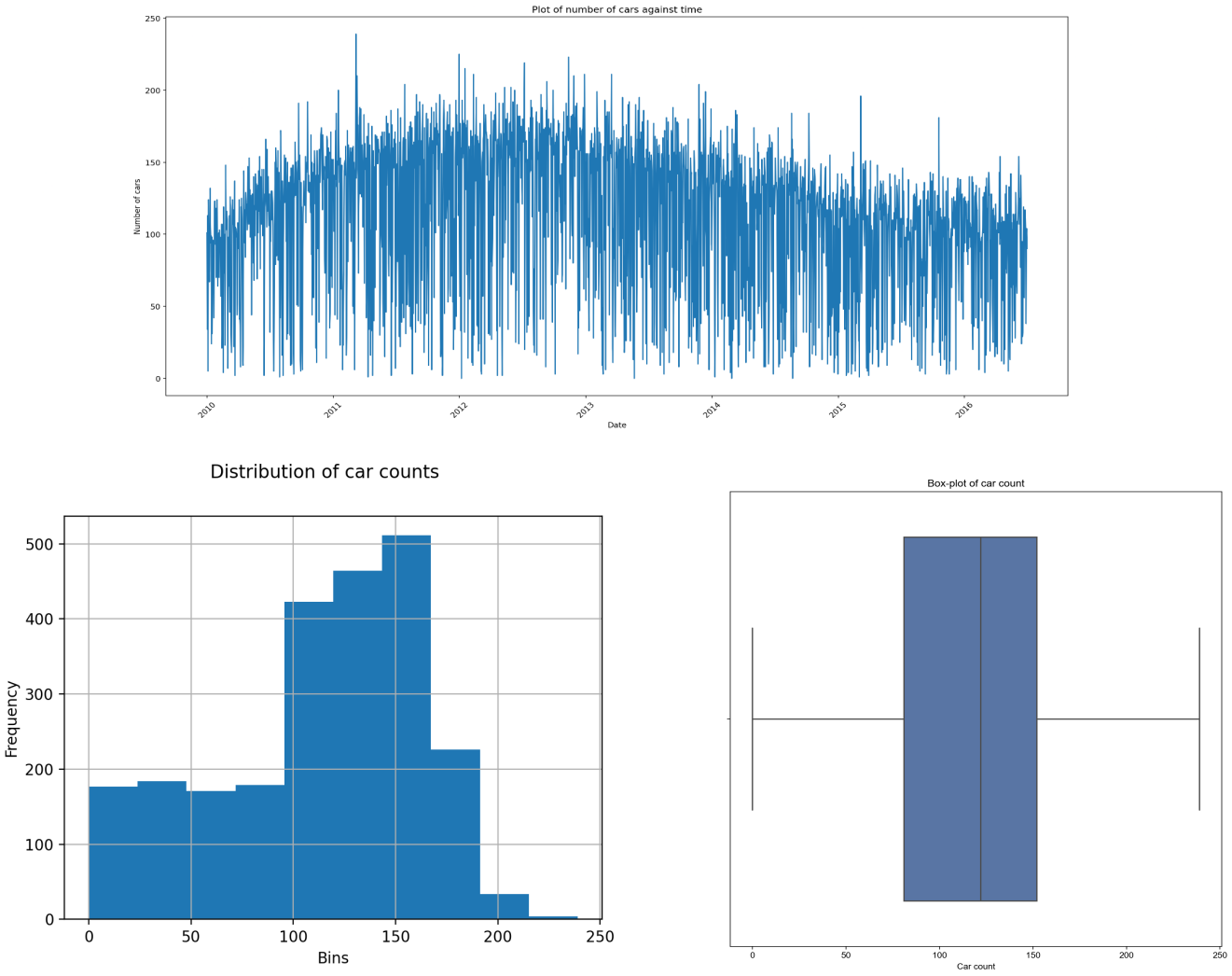


Figure 1. (a) Plot of car count against time; (b) histogram; (c) box-plot of car counts

From the plots in figure 1, we observe that

- the number of cars in the parking lot increased between 2010 and mid-2012, decreased thereafter until 2016 and then possibly increased from there onwards;
- there are some days when there are no cars were counted in the parking lot;
- the distribution of car counts is right skewed; and
- the data of car counts is noisy.

B Relationship of car counts with cloud indicator variable

We begin examining the relationship between car count and cloud indicator by plotting the car counts against time on cloudy and clear days. From figure 2, since the mean of (a) is lower than that of (b), we observe that less cars are counted on average on cloudy days than on clear days. Cloudy days also have higher variance in car counts than clear days, and contribute more to the noisiness in the overall car count data than clear days do. We also observe that the trends of increase and decrease in car counts, discussed in section A above, is attributed to the pattern of car counts on clear days, and not on cloudy days.

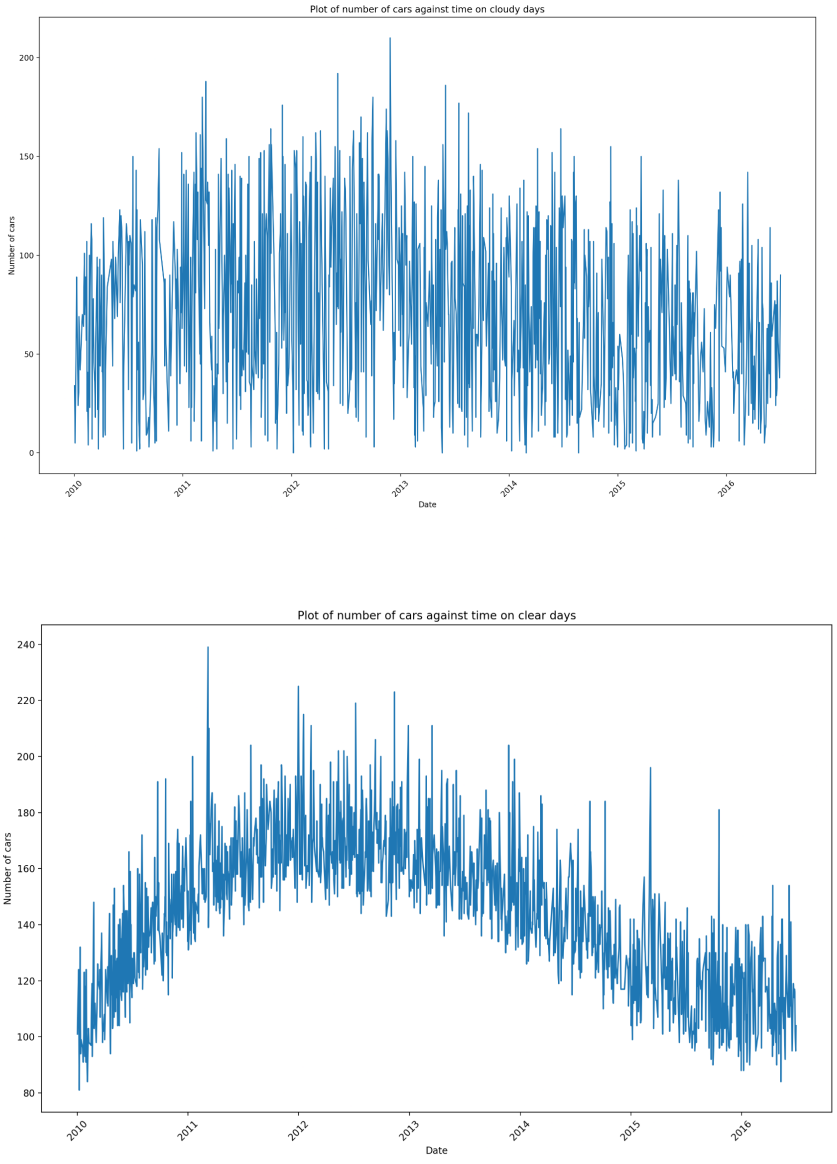


Figure 2. Plot of cars against time on (a) cloudy days and (b) on clear days.

The values of mean and standard deviations, and the box-plot on clear and cloudy days (figure 3 below), confirms the observations of mean and variance made above. We also note that the instances of no cars being counted on the parking lot occurs only on cloudy days.

```
(Pdb) data.groupby("cloud_indicator").mean()['car_count']
cloud_indicator
clear     143.811530
cloudy     71.602941
Name: car_count, dtype: float64
```

```
(Pdb) data.groupby("cloud_indicator").std()['car_count']
cloud_indicator
clear     26.038077
cloudy    44.428109
Name: car_count, dtype: float64
```

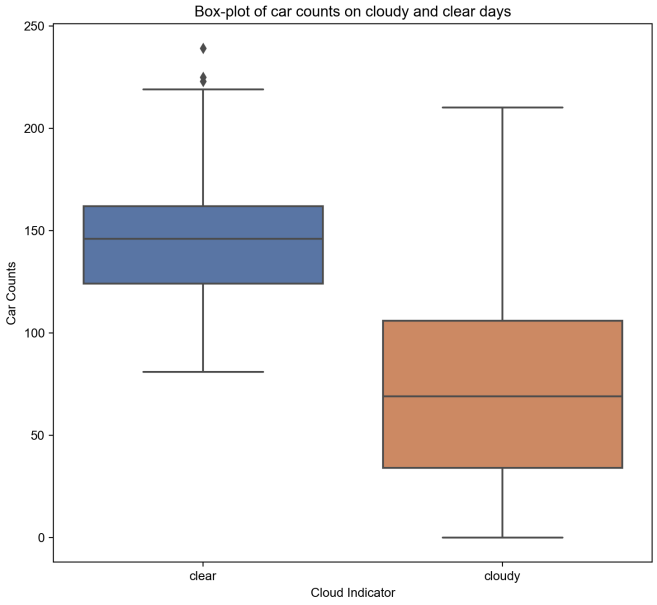


Figure 3. (a) Mean and standard deviation of car counts, and (b) box plot of car counts on clear and cloudy days.

The histogram in figure 4 shows disparate distributions of car counts on clear and cloudy days, and that there are more clear days (57%) than cloudy days in the data set

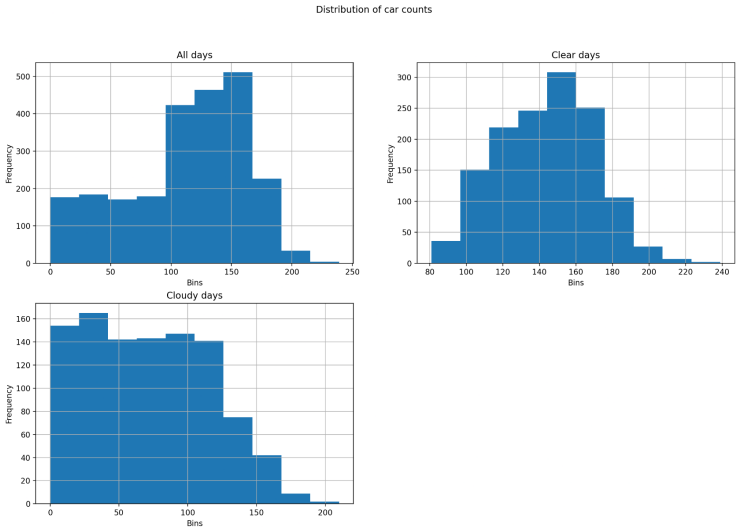


Figure 4. Histogram of car counts on (a) all days, (b) clear days, and (c) cloudy days.

We posit that the number of cars counted on cloudy days is significantly less than on clear days. We establish this relationship using the independent t-test, with the null hypothesis being that the means of cars on clear and cloudy days are not different. We reject the null hypothesis with 95% significance level since the t-statistics for difference of means was 49.55 (p-value = 0.00).

This relationship could be due to clouds causing poor visibility for satellite cameras, leading to unclear images being taken by them on cloudy days. As a result fewer cars are identified by the computer vision algorithms on cloudy days.

We analyze the consistency of the data on cloudy and clear days to examine its reliability. We do this using the coefficient of variation (where CV = (Std. Dev / Mean) * 100), where a high CV implies low consistency. We compute the CV for clear, cloudy and overall days as shown in figure 5.

Coefficient of variation on clear, cloudy and overall days

	Clear	Cloudy	Overall
Coeff. of variation	18.10%	62.05%	44.45%

Figure 5. Coefficient of variation in car counts on all days,clear days, and cloudy days.

We see that data on cloudy days has a lower consistency (by approximately a factor of 3.5), and hence is less reliable, than that on clear days. The overall data also has a low reliability.

C Dependence of car counts with time and trend analysis

We establish the dependence of car counts with time at daily, weekly and monthly timescales by examining the nature of the underlying process in the series at these timescales. The data set is kept as is for daily time series, and is summed at weekly and monthly intervals to create the weekly and monthly time series respectively. We then plot the autocorrelation values (the ACF or correlogram) and the partial autocorrelation values (PACF) for the series at these time scales. We observe that the series do not exhibit a moving average (MA) process at any timescales since the ACF plots in figure 6 do not drop (i.e. cuts-off) to 0 after any lag. This rules out stationarity, hence time independence, due to the MA process. The ACF and PACF plots are symptomatic of the series exhibiting flavors of auto regressive (AR) or auto regressive moving average (ARMA) processes. Furthermore, the ACF at weekly and monthly timescales are atypical of process being non-stationary since the coefficients slowly decay to zero.

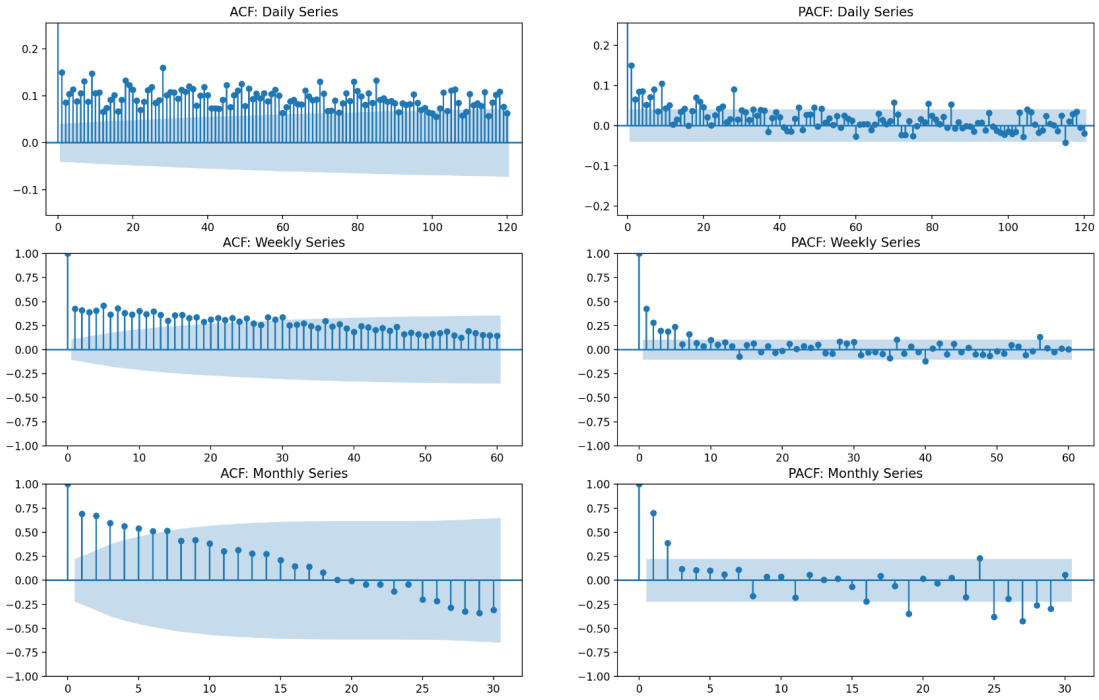


Figure 6. (a) ACF and PACF plots at daily, weekly and monthly timescales

Finally, the correlation coefficients are significant at all the timescales, albeit small in the series of daily timescale. Therefore the car count data has significant dependence with time on all time scales but more strongly on weekly and monthly timescales.

Trend analysis

We examine the existence of trends at days of the week and months of the year. We compute a 30-day rolling mean to derive trends from the data of days of the week since the data was fairly

noisy. From figure 7a, we observe that trends across different days of the week are similar to each other. Furthermore, the trends across different months of the year are also similar to each other as seen in figure 7b.

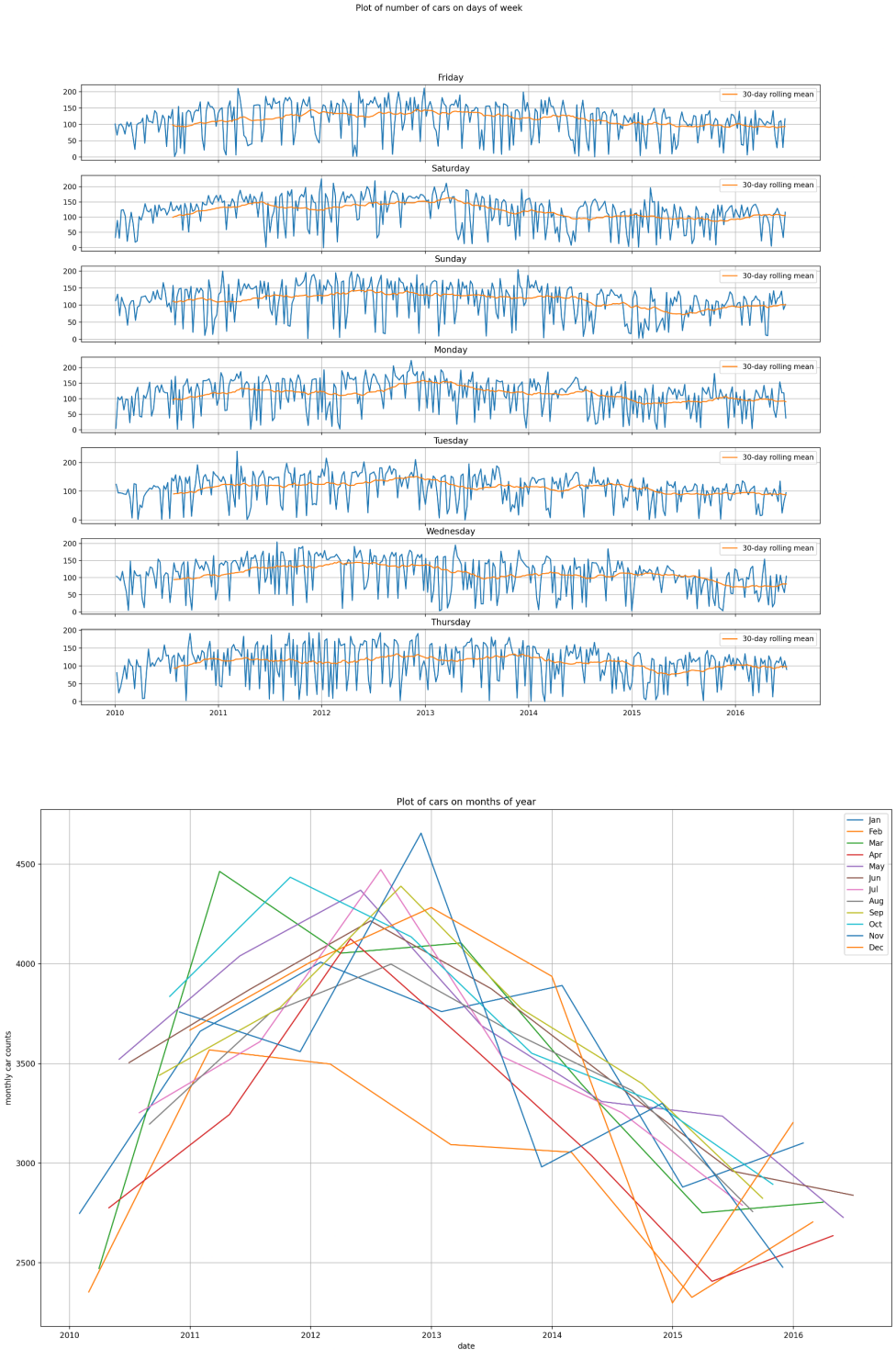


Figure 7. (a) Trends at days of the week (b) months of the year

D Relationship of car counts on weather

We begin analyzing the relationship of car counts on weather by first plotting the two variables and visually observing that there is lack of evidence of any form of relationship between them, as shown in figure 8.

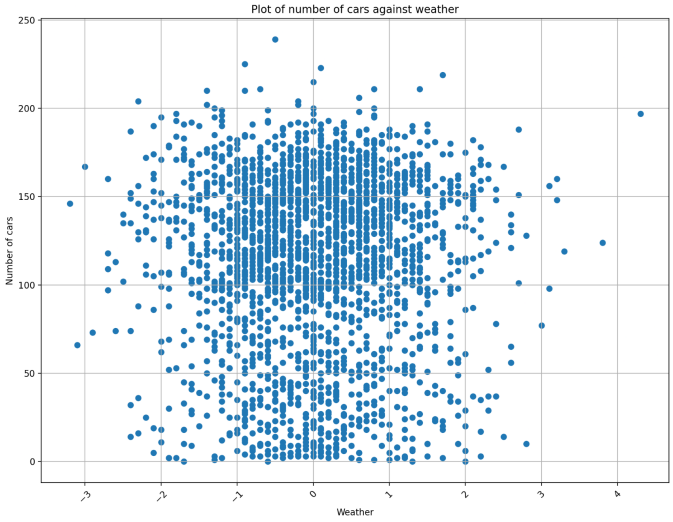


Figure 8. Scatter plot of the number of cars against the weather variable.

We then examine the autocorrelation of the weather series at various lags using the ACF and PACF plot, and observe that the correlation values are significantly close to 0 at all lags, shown in figure 9.

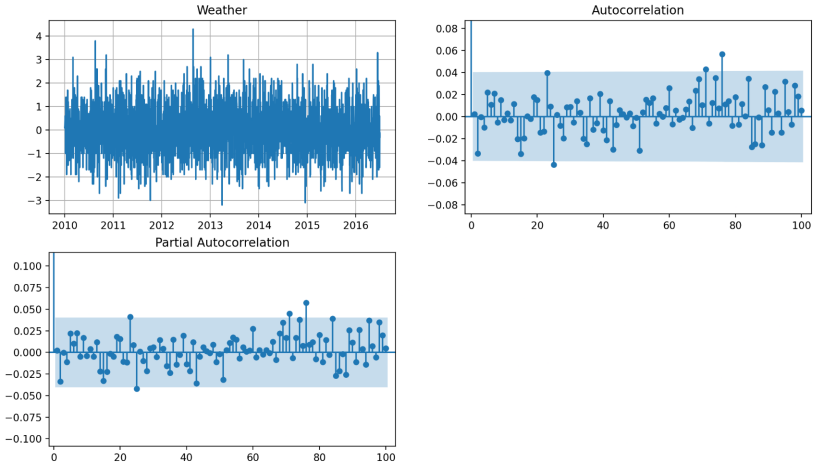


Figure 9. Time, ACF and PACF plots of the weather variable.

Finally we observe that the weather variable has zero-mean and unit-variance. This shows that weather is white noise and has no significant relationship with the number of cars.

E Influence on working with this data

From analysis in section B we infer that the cloud indicator variable will play an important role in the forecasting model. This is because we retain cloudy data in our dataset for forecasting car counts on cloudy days and the cloud indicator variable will aid in distinguishing between these two conditions. The reason for retaining cloudy data is that whilst the data on cloudy days is unreliable, it is still fairly valid (i.e. accurate) since it depicts trends that are fairly similar to the trends of car counts on clear days. This can be seen in figure 10 below, where the trend of car counts on cloudy days increases and decreases in the similar time periods as the trends on data for clear days does.

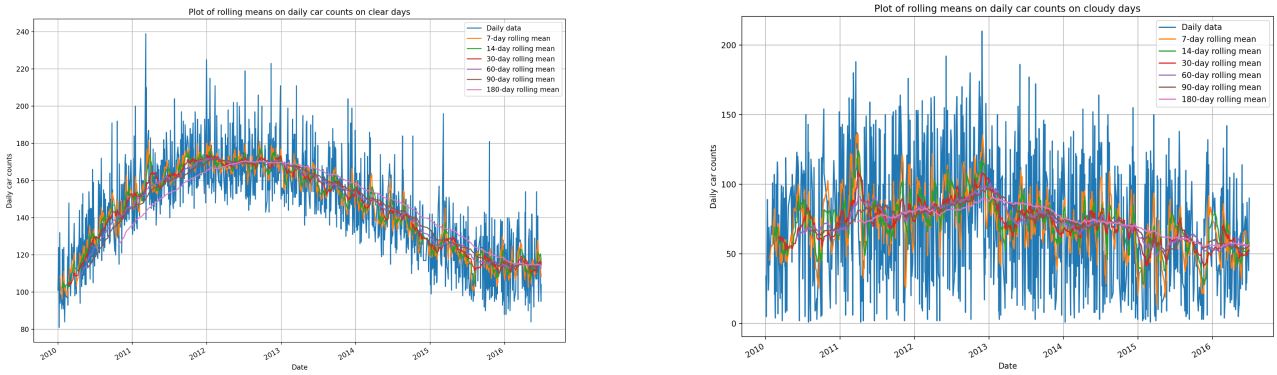


Figure 10. Trends of car count on (a) clear days and (b) cloudy days .

From analysis in section C, we observe that the car count data has significant time dependence, i.e. is non-stationary, and so will require pre-processing to make it stationary in order to build a model for forecasting purposes. We also note that days of the week and months of the year do not hold a lot of information in terms of explaining car count since trends within them are similar to each other.

The analysis in section D informs us that the weather variable will not be of much use in building a forecasting model.

FORECASTING

We propose the use of an autoregressive integrated moving average with an exogenous variable (ARIMAX) model to forecast car traffic one day in advance. We use the data at daily timescales and include the cloud indicator variable as an exogenous regressor. The details of the model are described henceforth.

A. Parameters of the model

We will include the cloud indicator variable since we demonstrated in section B above that it had a significant relationship with car count. We exclude the weather variable since we showed that it was white noise and hence did not have a relationship with car count.
We examine the variable day of the week for having a significant relationship with car count. We begin by observing that the mean number of cars are similar on different days of the week (figure 11).

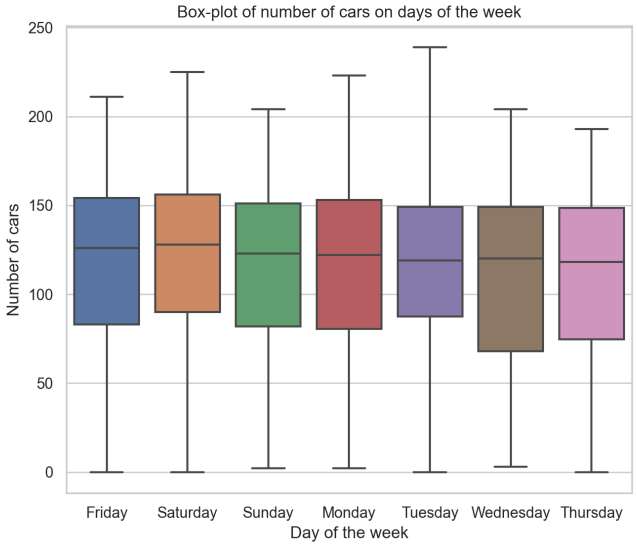


Figure 11. Box-plot of car count on days of week.

This indicates that the variable day of the week does not significantly affect the car count. The analysis in section C above corroborates this conclusion. Therefore the variable day of the week is not included in the model.

Guard against over-fitting

In selecting the appropriate model, we avoid picking a model that overfits by examining the Akaike and Bayesian Information criterions (AIC and BIC). These values weigh the risks of over- and under- fitting the data. We choose a model that lowers their values.

Treatment of cloudy data

We include car counts on cloudy days in our model, and identify it using the cloud indicator variable.This variable is included as an exogenous variable in the ARIMA model. Given its binary nature, it is a dummy variable and its value is indicative of the amount of offset the model will apply in the number of cars between clear and cloudy days.

Deducing the order of ARIMA

We examine the ACF and PACF plots of the car counts data at zero, first and second order differences in efforts to deduce the order of the ARIMA model as shown in figure 12.

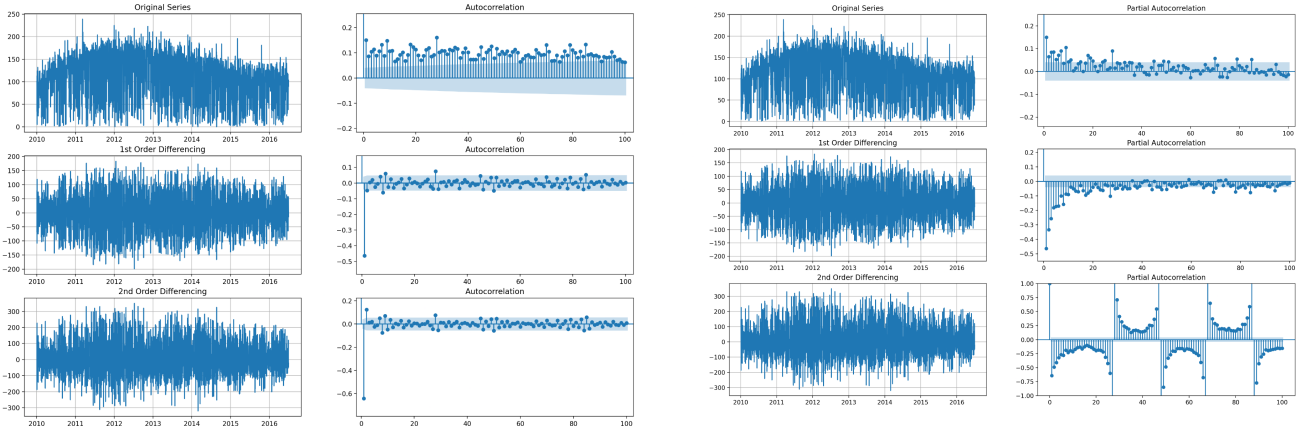


Figure 12. (a) ACF and (b) PACF plots of car count series at zero, first and second order differences.

We observe that the series would be better modelled at first-order difference, i.e. I(1), where the plots show that the series become sufficiently stationary. Furthermore, we note that the series is over-differenced at second-order since adding an additional difference causes the auto-correlation value at lag-1 to exceed -0.5. From the ACF plot, we observe that the first-order differenced series has strong MA(1) tendencies and, since the PACF shows slow decay, an ARMA (1,1,1) model would also be a good contender for this series. Note: Higher order ARMA models may provide a slightly better in-sample fit but may be susceptible to higher out-of-sample errors due to over-fitting tendencies.

We corroborate our deduction using the *auto_arima* method of the library *pmdarima* in Python which selects an appropriate model for the series; the method takes AIC into consideration to guard against selecting models that overfit. Figure 13 shows the models tried by the *auto_arima* algorithm, together with their respective AIC values.

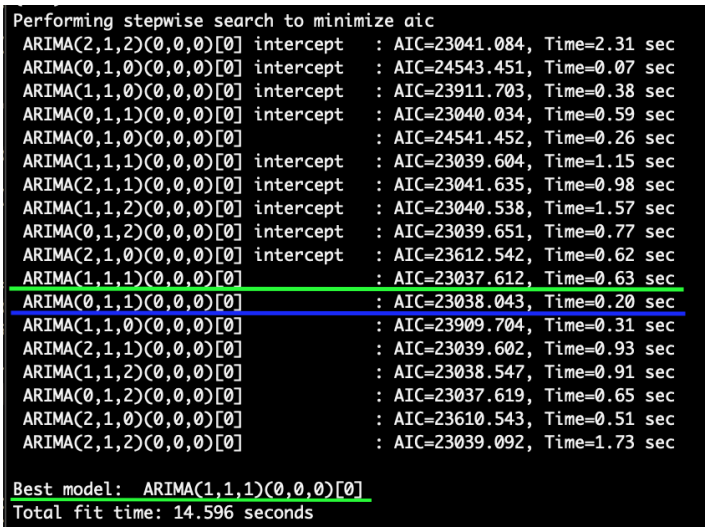
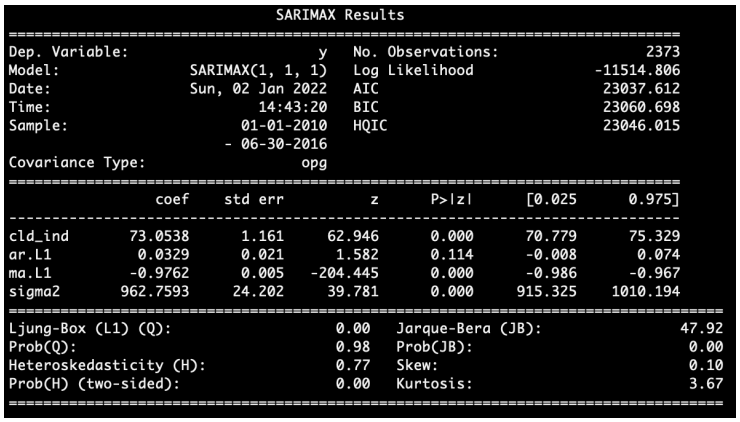


Figure 13. ARIMA models tried by the *auto_arima* method in python

We note that the ARIMA (1,1,1) model is selected since it has the lowest AIC values. A simpler, ARIMA (0,1,1) comes in as a strong contender and yields lower BIC values (shown in figure 14b). The results of fitting the models are shown in figure 14. We observe that the value of the AR coefficient in the ARIMA(1,1,1) is very small and not significant as shown in figure 14a. This supports our speculation that the ARIMA(0,1,1) model may have a better generalization error.

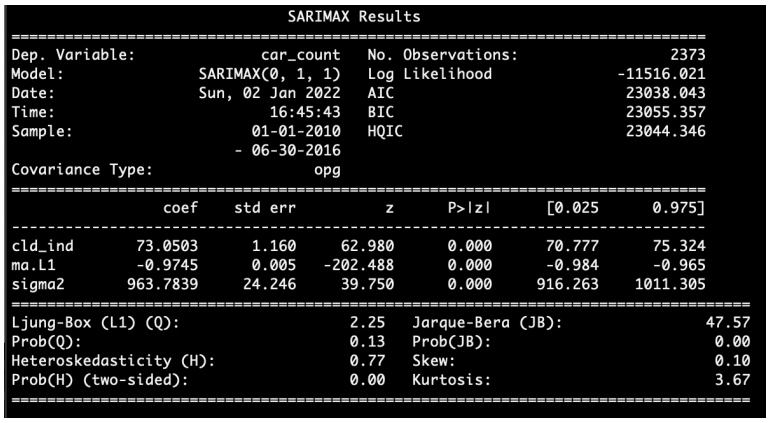


SARIMAX Results

Dep. Variable:	y	No. Observations:	2373
Model:	SARIMAX(1, 1, 1)	Log Likelihood	-11514.806
Date:	Sun, 02 Jan 2022	AIC	23037.612
Time:	14:43:20	BIC	23060.698
Sample:	01-01-2010 - 06-30-2016	HQIC	23046.015
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
cld_ind	73.0538	1.161	62.946	0.000	70.779	75.329
ar.L1	0.0329	0.021	1.582	0.114	-0.008	0.074
ma.L1	-0.9762	0.005	-204.445	0.000	-0.986	-0.967
sigma2	962.7593	24.202	39.781	0.000	915.325	1010.194

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	47.92
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	0.77	Skew:	0.10
Prob(H) (two-sided):	0.00	Kurtosis:	3.67



SARIMAX Results

Dep. Variable:	car_count	No. Observations:	2373
Model:	SARIMAX(0, 1, 1)	Log Likelihood	-11516.021
Date:	Sun, 02 Jan 2022	AIC	23038.043
Time:	16:45:43	BIC	23055.357
Sample:	01-01-2010 - 06-30-2016	HQIC	23044.346
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
cld_ind	73.0503	1.160	62.980	0.000	70.777	75.324
ma.L1	-0.9745	0.005	-202.488	0.000	-0.984	-0.965
sigma2	963.7839	24.246	39.750	0.000	916.263	1011.305

Ljung-Box (L1) (Q):	2.25	Jarque-Bera (JB):	47.57
Prob(Q):	0.13	Prob(JB):	0.00
Heteroskedasticity (H):	0.77	Skew:	0.10
Prob(H) (two-sided):	0.00	Kurtosis:	3.67

Figure 14. Results of fitting (a) ARIMA(1,1,1) abd (b) ARIMA(0,1,1) models

We examine the residuals yielded by these models in section B below.

B. Residuals of the model

We analyze the residuals of the fit from both the ARIMAX(1,1,1) and ARIMAX(0,1,1) models to see if they are approximately white noise. To do this we begin by first plotting standardized residuals against time and visually observe that they look like white noise for both models. We then plot their distribution and compare it with the Gaussian distribution, as well as generate a Q-Q plot against a gaussian distributed variable. These two plots show that the residuals from both the models are approximately gaussian distributed. We then test for stationarity by plotting the ACF plot and observe that the residuals for the models are serially uncorrelated since the autocorrelation values are very small at all lags. This is confirmed by the Ljung-Box test in figure 14. Therefore the residuals are white which tells us that the coefficients of the model adequately captures the true signal that is generating the data. The plots in figure 15 show the aforementioned plots.

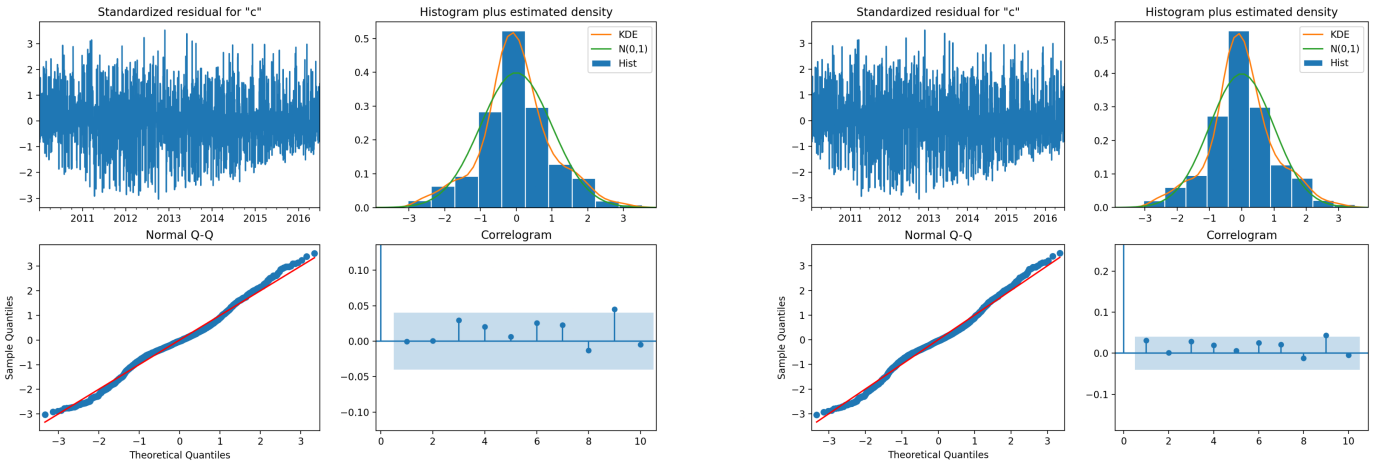
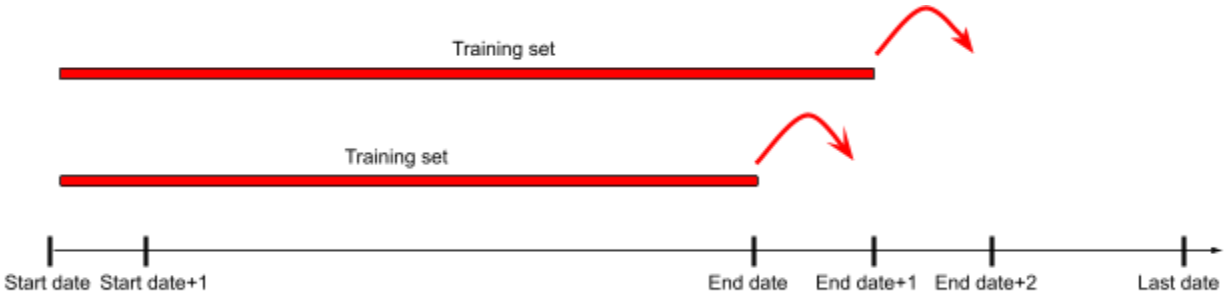


Figure 15. Residuals of fitting the (a) ARIMA(1,1,1) abd (b) ARIMA(0,1,1) models.

C. Out of sample test

We establish the out of sample performance of the two ARIMAX models by training them on data points of car count and cloud indicator from the starting date (say 'start date') of data up to a certain date in time ('end date'), and using the trained model to predict the number of cars on the following date ('end date + 1 day''). We then append the training set by one date, by including the true value of the data point (and not the value predicted by the model) from the following day ('end date + 1') into our training set and predicting the value for the day after that (see diagram below).



This form of out of sample testing algorithm is provided by the *get_prediction* method in the *stats model_ARIMA* library. The method of training on historical data points and testing on future data points prevents data leakages where there are data points from future dates entering the training set leading to bias in estimation of true out of sample performance.
We then compute the root mean square and mean absolute errors between the predicted and true car count values.
The out of sample results of the two models are shown in figure 16 below. See observe that the ARIMA(1,1,1) model performs slightly better than the ARIMA(0,1,1) model, supporting the best model chosen by the *auto_arima* method based on AIC values (see section A above).

Out of sample performance of ARIMA models

	ARIMA(1,1,1)			**ARIMA(0,1,1)**		
	Clear days	Cloudy days	Total	Clear days	Cloudy days	Total
RMSE	15.97	39.39	28.63	16.00	39.43	28.67
MAE	13.07	33.55	21.98	13.08	33.62	22.02

Figure 16. Out of sample performance of the ARIMA(1,1,1) and ARIMA(0,1,1) models

We see that the model has a higher error in predicting car count on cloudy days than on clear days. This is expected given higher levels of unreliability in the cloudy data.

D. Generating uncertainties on the forecasted value

The ARIMAX model is implemented by the Kalman filter algorithm. This state space method computes an estimate of the state and a level of uncertainty in the form of an error covariance matrix. This is used to generate uncertainties around the forecasted values and establish desired levels of confidence around it. Figure 17 shows the actual car counts (in blue) and the one step ahead forecasted car counts (in orange). The 95% confidence interval bounds on the prediction are shown in grey.

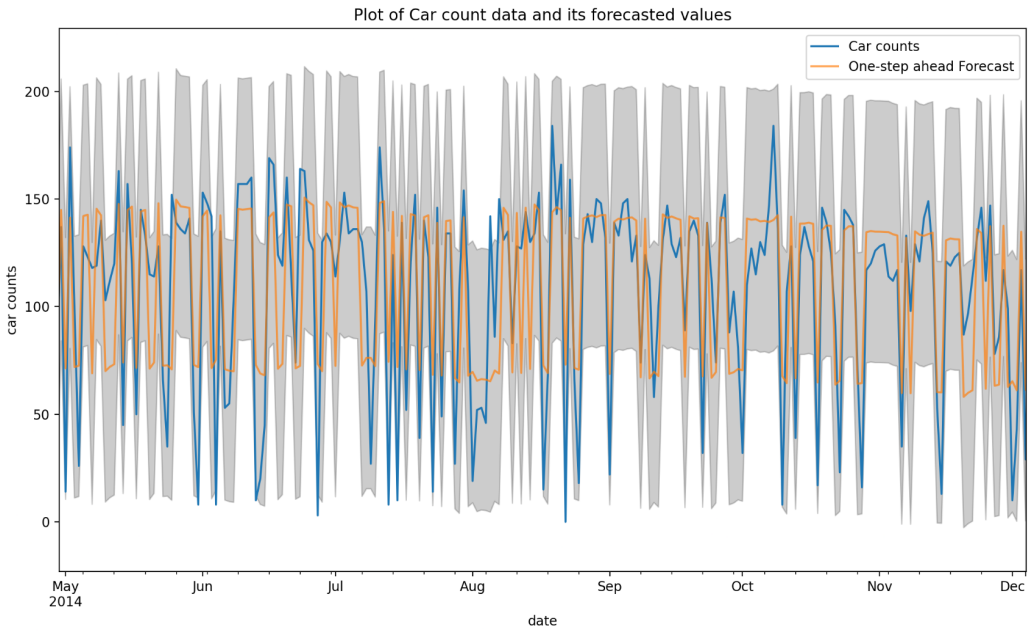


Figure 17. Plot of actual and forecasted car counts across time