

Intro to ML Final Exam 2

Abhinav Sharma

1/8/2021

Raw .Rmd code available at — <https://github.com/abhinav-sharma-6167/Intro-to-ML-2>

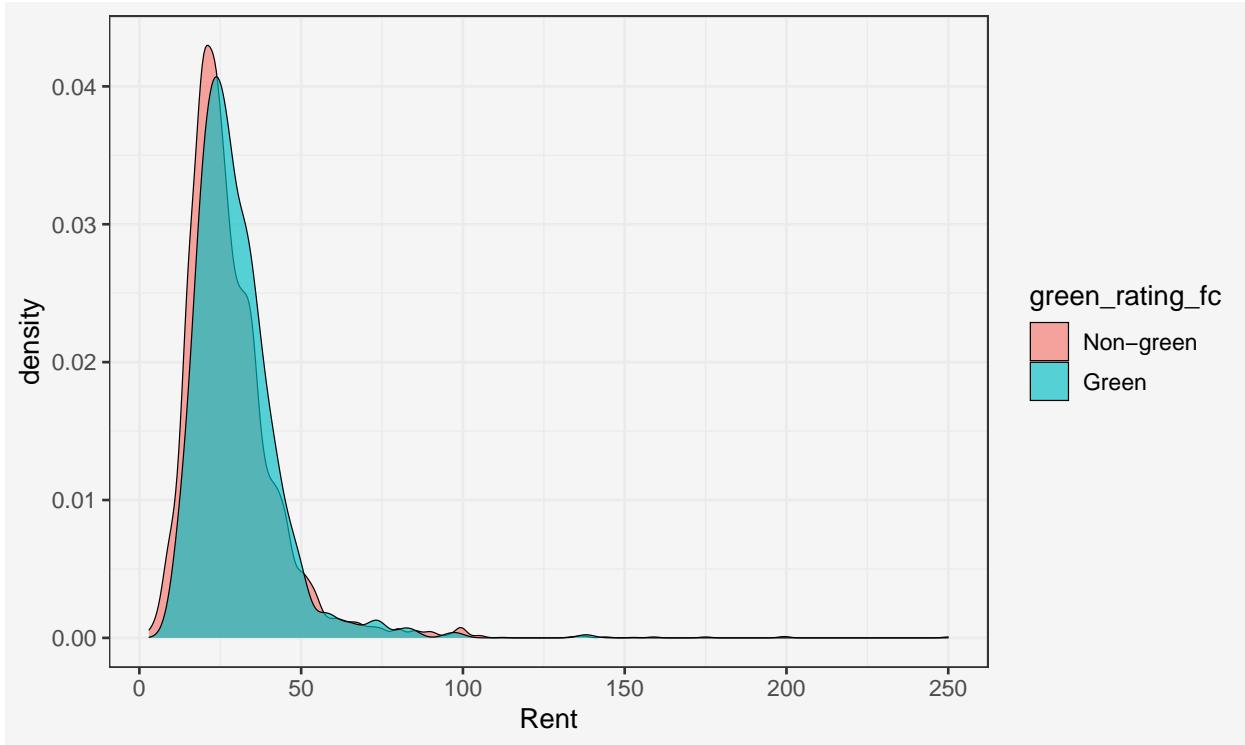
Setup - Loading libraries and setting working directory. Key libraries include ggplot2, plotly for exploration; data.table and dplyr data wrangling. Setting up custom ggplot theme for all plotting purposes.

Visual story telling part 1: green buildings

The details given in the case helps us get a sense of the problem. To quickly summarize, the data encompasses of 7,894 commercial rental properties of which 685 are green buildings. To avail a control set for the 685 properties, the creators of this data leveraged all non-rated buildings within a quarter-mile radius of the green-certified building. On average we have 12 non-rated nearby properties for each green property. The idea is green houses would be more attractive living options given lower recurring costs, better indoor environments, longer economically valuable lives and in general, the good PR they enjoy. The goal is to validate whether investing in a green building be worth it, from an economic perspective. Specifically, in a new 15-story mixed-use building on East Cesar Chavez, just across I-35 from downtown with baseline construction costs being \$100 million and a 5% expected premium for green certification.

Building upon this premise, we now start exploring the data. We start by checking the number of green houses and distributions of the variables used in stat-guru's analysis to stress-test the assumptions made.

Percentage of green buildings :
8.68 %



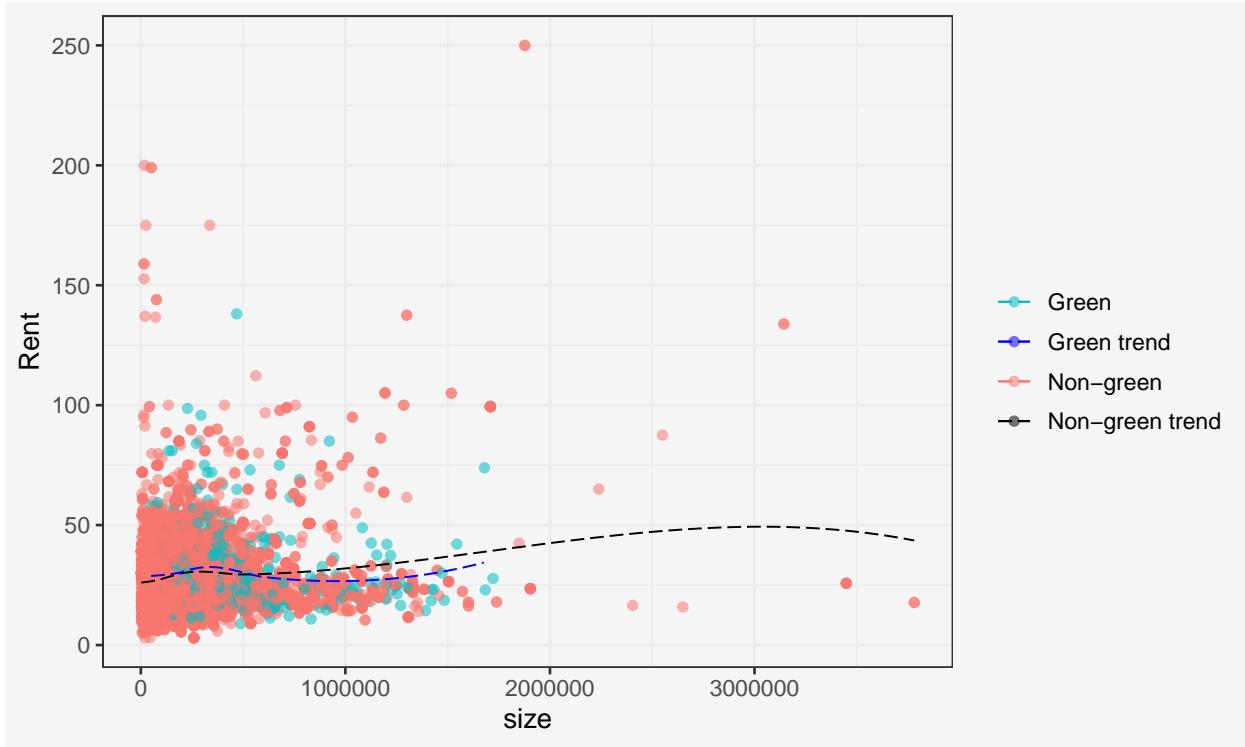
Summary Stats :

	green_rating	Mean_Rent	Med_Rent	SD_Rent	IQR_Rent
1:	1	30.03	27.60	12.96	14.04
2:	0	28.44	25.03	15.33	14.75

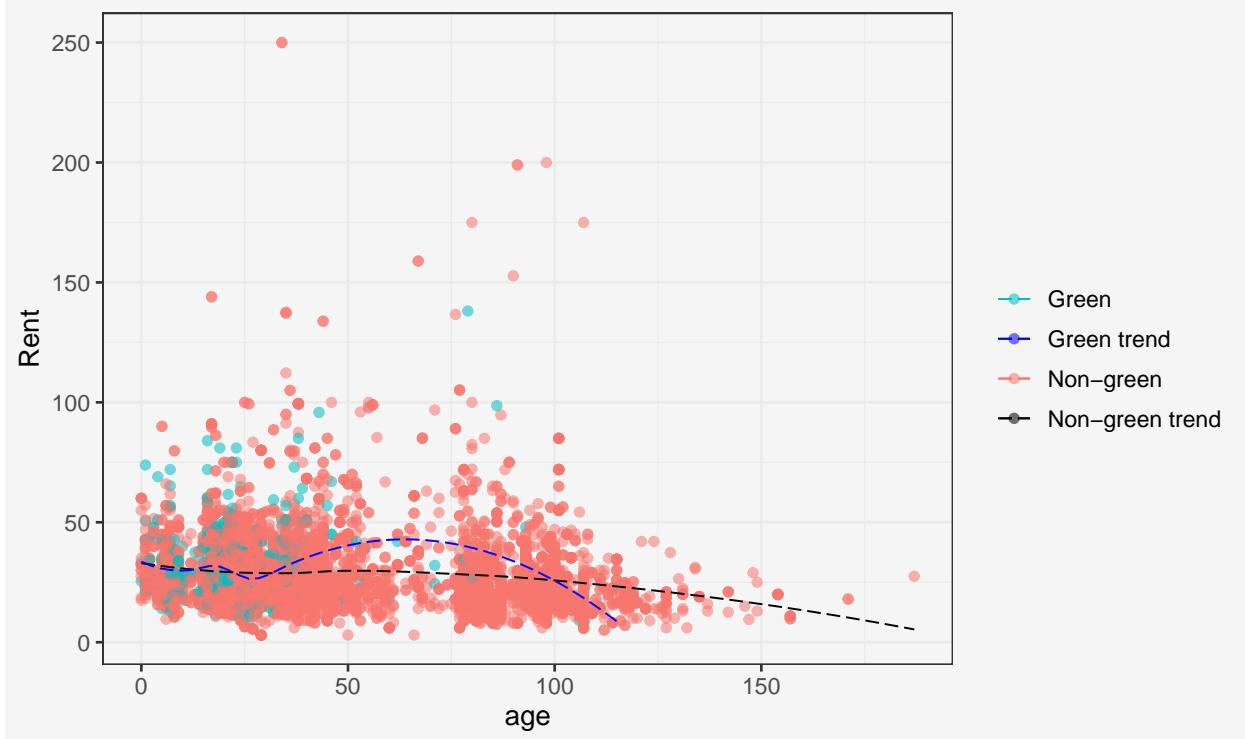
We see Rent for green houses is slightly larger than other houses. The rent variable is extremely right skewed so it makes sense to use median as the measure for centrality instead of mean. The statistical summary suggests Rent distribution is slightly more spread out for non-green buildings given it has higher SD and IQR. However, the graph looks pretty much the same for both categories implying, non-green building Rent has more outliers compared to green buildings.

However, this difference in medians could arise due to multiple confounding variables. Few hypotheses for this price change could be : 1. Properties with higher size may have higher rent 2. Older houses might have lower rent 3. Renovated houses may have higher rent 4. Houses in better class society and with more amenities may have higher rent 5. Houses with more Gas and Electricity costs may have higher rent, given they have more high energy-consuming facilities 6. Only specific type of green buildings (say LEED) may offer difference in rent

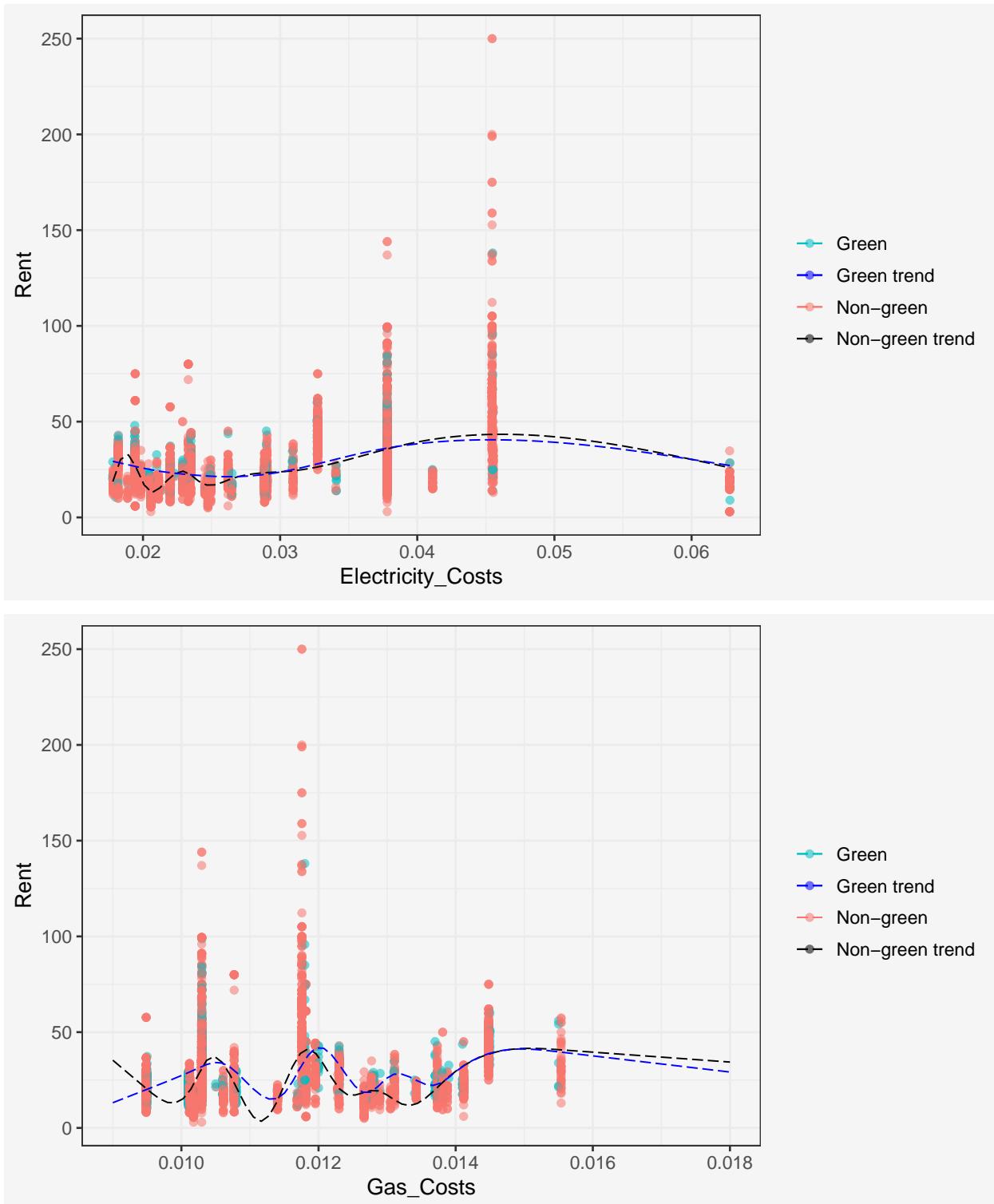
Let's explore them one by one. We use the non-parametric Generalized Additive Model (GAM) based smoothening to capture the non-linearity in the trends in our scatter plots.



We observe almost no difference in Rent for green buildings vs other until size 900000, post which rent of non-green buildings slightly increases. However, this difference could be due to more outliers for non-green category and the confidence intervals of the fitted curves would intersect, negating any significance associated with this difference.



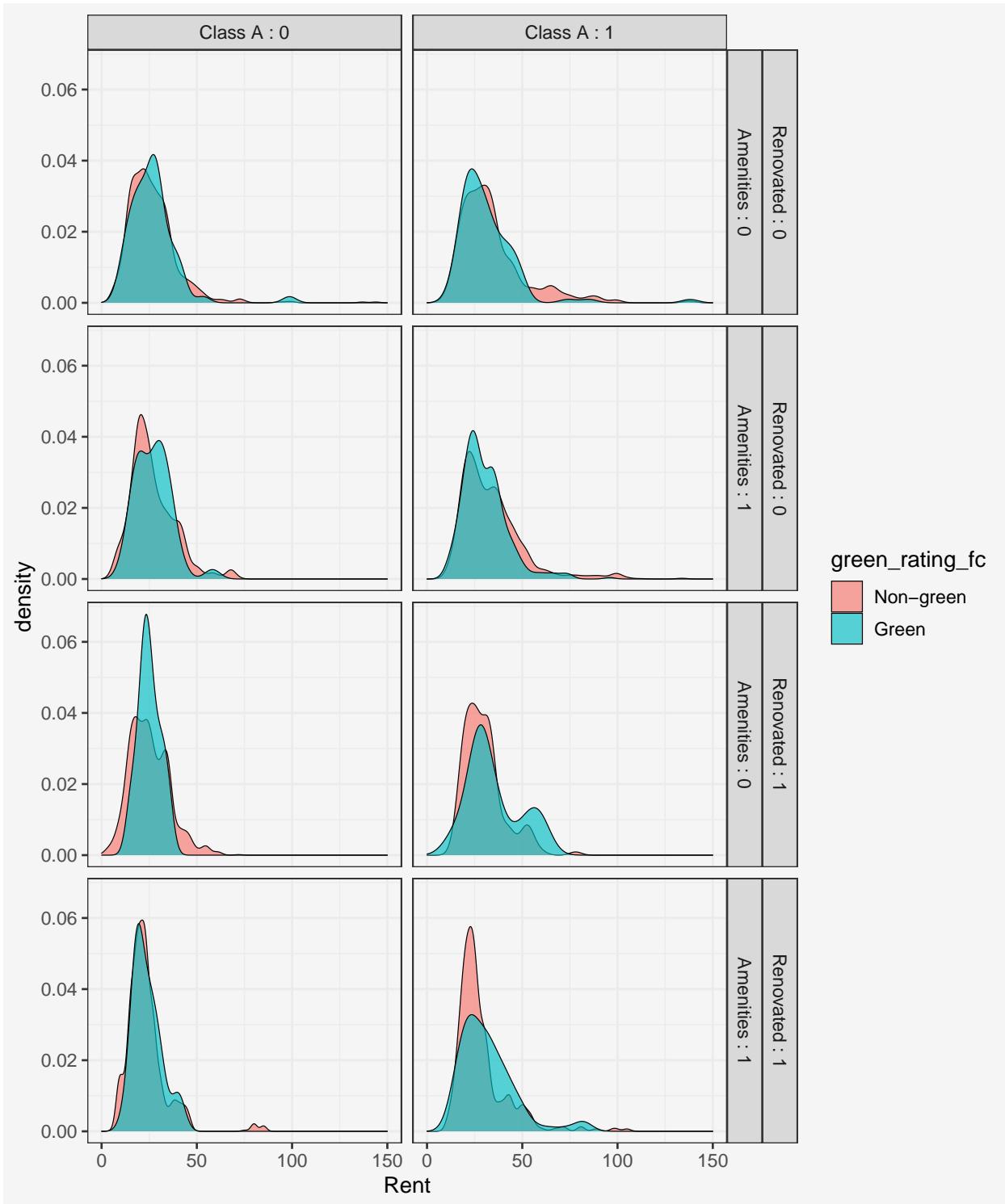
We see for houses aged 40-90 have some differentiation in rent with respect to houses being green vs non-green



There's hardly any difference in Gas and electricity utilization, apart from a small segment of Gas costs between 0.011-0.012 where green buildings have lesser gas costs.

Even with continuous variables, we get the sense that the difference between prices of green building vs not wouldn't be as straightforward as difference of median rents of the two categories. The localized regression trendlines convey that for certain properties of a house, some of the price difference could be explained by confounding variables such as age, gas costs.

We also explore categorical variables and check whether different categories lead to change in Rent.



Rent of house being green vs non-green differs widely in distribution, specifically in cases where house is renovated, is in upper class society and has additional amenities. Given, East Cesar Chavez looks very developed and say, the house has undergone renovations and includes amenities, the stat-guru's assumption of all green houses having incremental rent of \$2.6 than their non-green houses is erroneous.

The median price difference in such a case would be : \$ 0.75

This implies that the duration of cost recuperation would be higher than 7.7 years. Assuming, the real difference in medians upon accounting for all confounding variables was \$ 0.75, the premium of investment in a green building would be recovered in about 26.67 years. Hence, availing green status to ensure more probitability might not be a good strategy.

Because the proportion of green buildings is very less, thereby green vs non-green buildings have unequal sample sizes. We can use Anova to have a final comparison of means of unequal samples and remove Rent higher than 100 as outlier. We can include covariates we've explored in the EDA above such as recurring costs, size and age of building, class and amenities to quantify their effect sizes.

	Feature	Coef	F_stat	P_val
1: green_rating_fc		6.5335	10.1345	0.0015
2: size		-0.0172	183.8325	0.0000
3: age		0.0000	56.7926	0.0000
4: renovated		0.0353	99.2252	0.0000
5: class_a		-2.9395	234.2568	0.0000
6: class_b		6.9193	39.3476	0.0000
7: amenities		2.7857	14.0708	0.0002
8: Gas_Costs		0.6698	5.0544	0.0246
9: Electricity_Costs		-692.6059	1612.2222	0.0000
10: empl_gr		756.3225	6.2707	0.0123
11: green_rating_fc:size		0.0354	4.5498	0.0330
12: green_rating_fc:age		0.0000	2.1967	0.1384
13: green_rating_fc:renovated		0.0493	1.8038	0.1793
14: green_rating_fc:class_a		1.9534	2.0126	0.1560
15: green_rating_fc:class_b		-0.2348	0.2521	0.6156
16: green_rating_fc:amenities		-2.7601	1.0341	0.3092
17: green_rating_fc:Gas_Costs		-1.1879	0.0308	0.8607
18: green_rating_fc:Electricity_Costs		40.3885	0.4524	0.5012
19: green_rating_fc:empl_gr		-30.2841	2.5275	0.1119
20: Residuals		0.0924	NA	NA

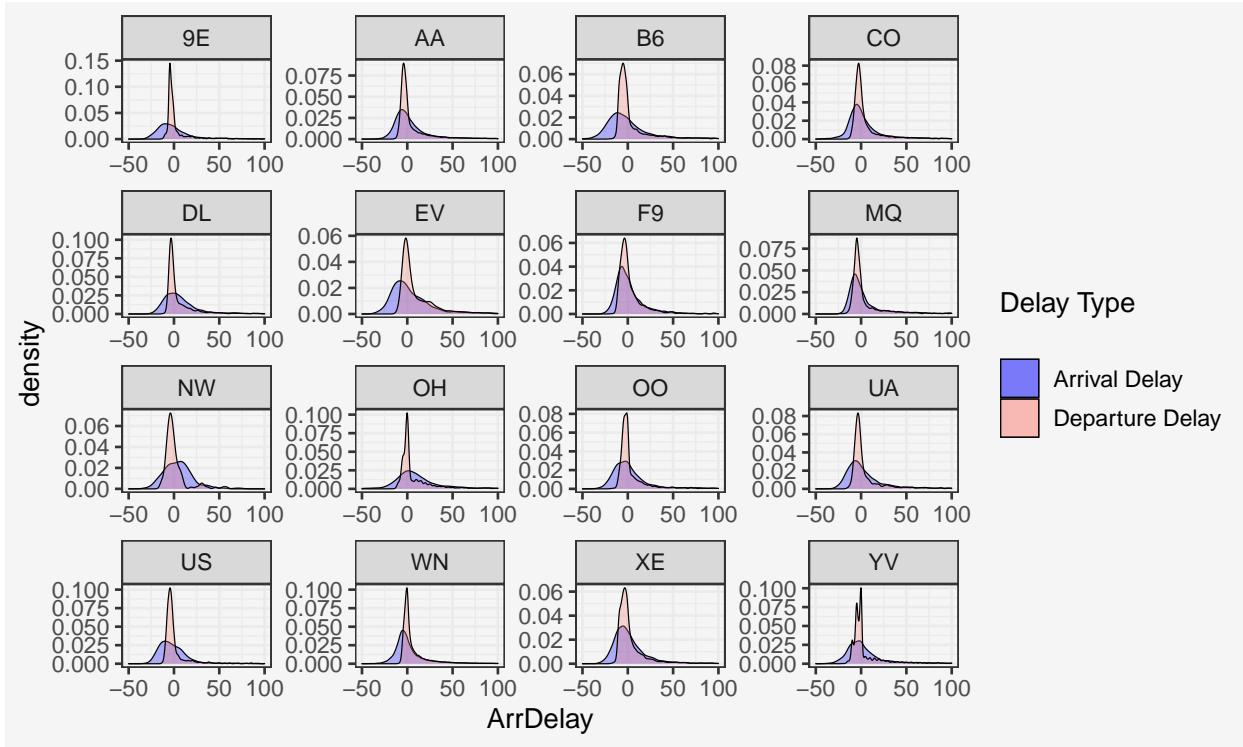
Supporting our EDA, we see all covariates have an impact on the rent of the building and solely attributing the rent to green status by differencing medians would be incorrect.

Visual story telling part 2: flights at ABIA

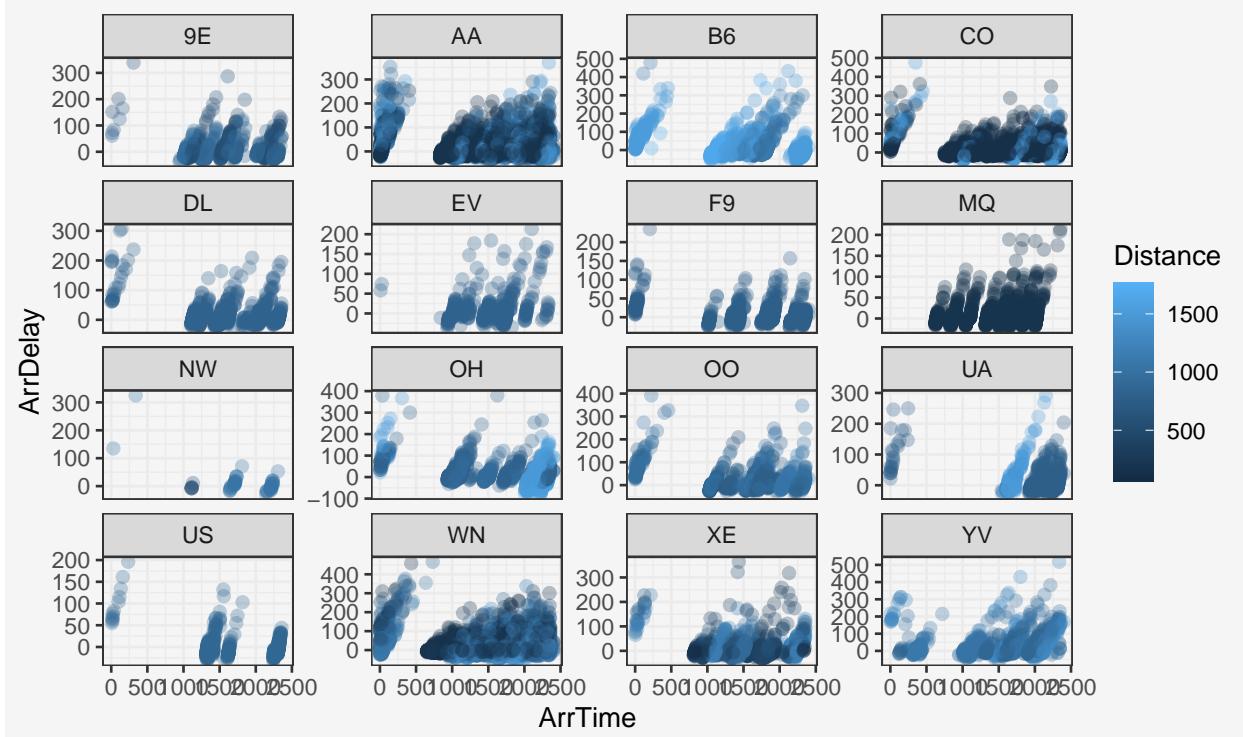
We start by exploring the data using summary functions such as str, summary, Hmisc::describe() and DescTools::Desc() to get a sense of the overall data.

Preliminary univariate EDA shows : 1. DepTime and ArrTime have 1.4% and 1.6% missing values respectively. Variables CarrierDelay, WeatherDelay, NASDelay, SecurityDelay and LateAircraftDelay all have 80.1% missing values. 2. There is some seasonality with respect to month with lesser flights towards year end in months Sept-Dec'08. Similarly there are fewer flights on weekends compared to weekdays. 3. Total 16 unique airline carriers with top three carriers operating most flights being Southwest (WN), American Airlines (AA) and Continental (CO) 4. ActualElapsedTime has bimodal distribution. ArrDelay, DepDelay distribution has very high kurtosis implying high number of outliers. Distance follows a spread-out distributions with 3 modes, meaning there could be some categorization such as local, national and international flights. Variables CarrierDelay, WeatherDelay, NASDelay, SecurityDelay and LateAircraftDelay 5. Most incoming flights to and from Austin are connected with the cities DAL, DFW, IAH, PHX etc.

We start by studying if there is any pattern in Arrival and Departure delay by airline carriers.

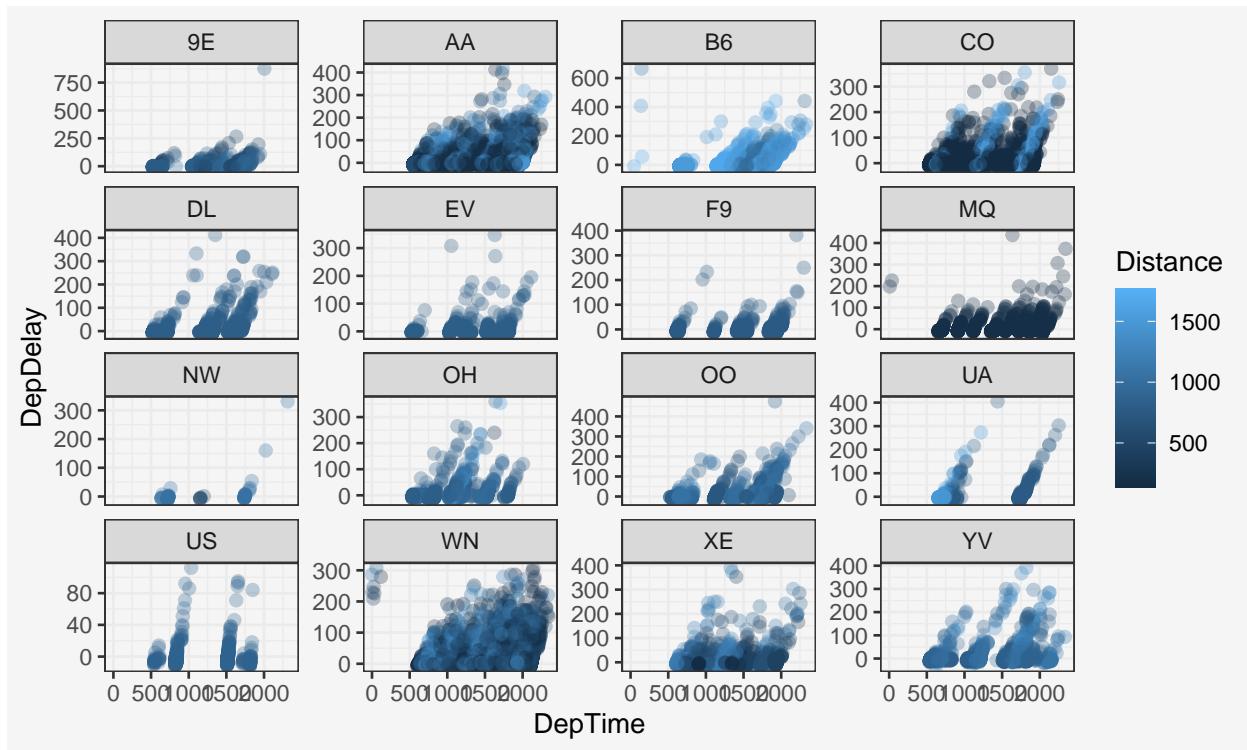


We observe that Departure delays for most carriers are often less in magnitude (be it positive or negative) than arrival delays, with an exception being F9 carrier where arrival vs departure delays are more or less the same.



Above graph shows Arrival Delays for flights arriving in Austin are particularly high when there are multiple flights of the same carrier arriving with very little time between each other. This could probably be explained by the air-space congestion issue. We also see marginally more delays when flights are for longer distances.

Carriers YV and B6 have higher delays associated with them.



We avail the insight into Departure delay for flights flying out from Austin. Times with multiple flights lined one after another are