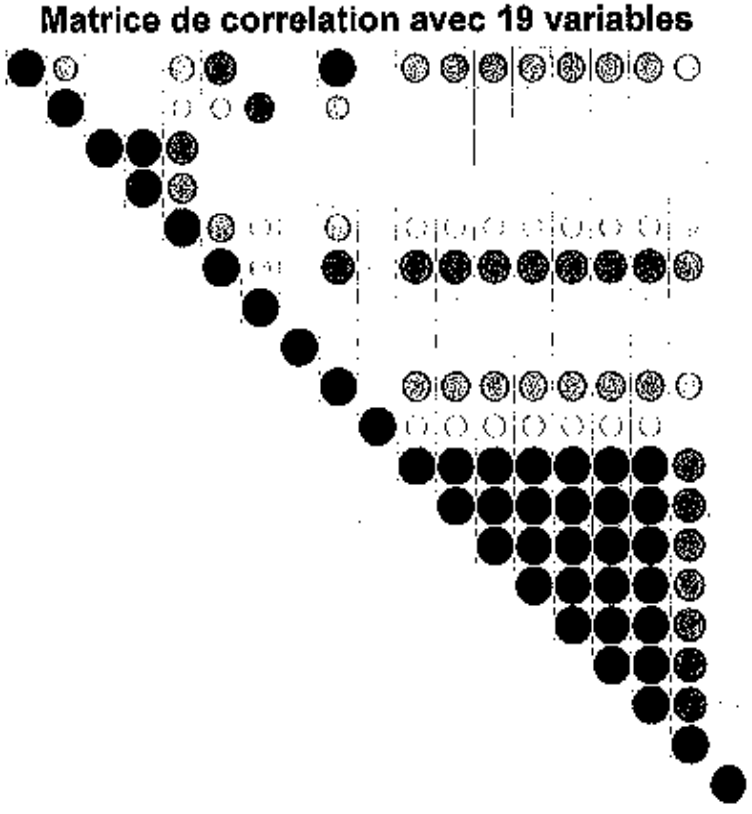
En tant que data scientits, nous allons devoir étudier l'ensemble de données à propos de l'immobilier d'appartements dans les quartiers résidentiels en Tehran, Iran. Nous allons donc expliquer deux variables qui sont : les coûts de construction et les prix de vente grâce à des modèles de régression. Pour cela, nous nous appuyerons sur 8 variables à propos des conditions physiques et financières du projets, de 19 variables économique sur plusieurs périodes de temps.

Dans un premier temps, nous allons étudié la variable des coûts de construction, appelée V9 dans le set de données.

On remarque dans un premier temps une forte corrélation avec de nombreuses variables, notamment V8, qui est la variable concernant le prix de l'unité au début du projet par m^2, ainsi que la deuxième variable à étudier qui est le prix de vente.

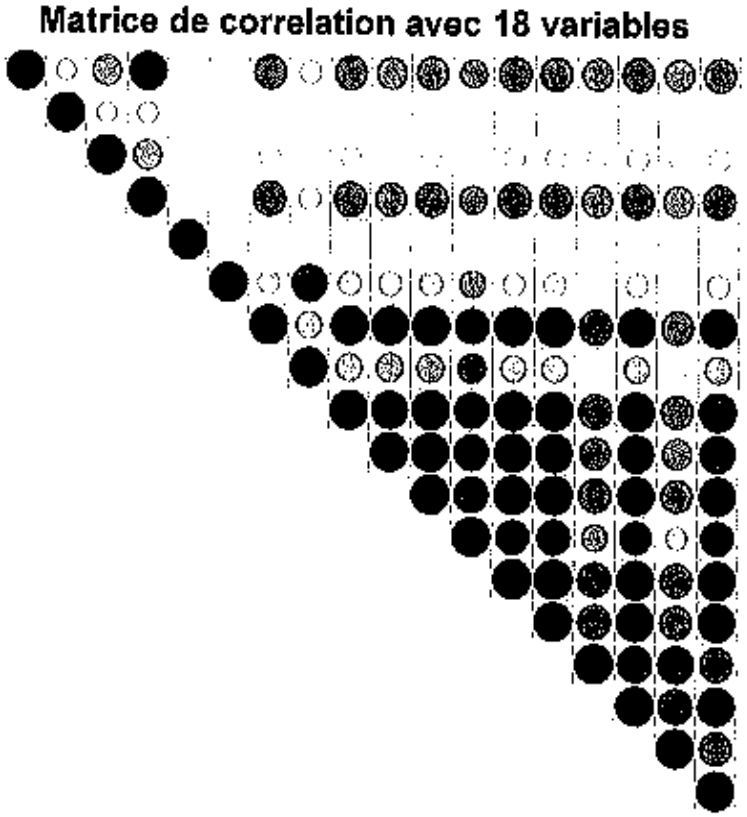
Lorsqu'on étudie un modèle linéraire utilisant nos 27 variables, on remarque que certains coefficients ne sont pas déterminés car ils sont trop similaire aux autres. Cela signifie qu'on prend en compte trop de variables. On va donc réduire le nombre de variables à ceux qui sont les plus significatifs.



Sur le graphe de corrélation ci-dessus, nouus avons réduit le nombre de variables de 27 à 19 afin de mieux visualiser le lien entre variables. Par souci de clarté, les noms de variables ont été omis. Mais grâce à la première ligne, nous pouvons effectivemeent confirmé une forte corrélation positive majoritairement entre la variable à étudier, soit les coûts de construction, avec nottament la variable V8, qui est le prix de mètre carré

au début du projet. on observe également une très forte corrélation entre les variables évolutives au cours du temps.

De la même manière, nous allons effectuer la même étude pour la deuxième variable sur le prix de vente.



De la même façon que la première variable, nous avons réduit le nombre de variables de 27 à 18. On peut alors comme avant remarquer avec la première ligne que le prix de vente est étroitement corrélé avec de nombreuses autres variables explicatives, et principalement avec la variable V5, qui correspond aux estimations de coute de construction. Evidemment on retrouve la forte corrélation des variables temporelles.

Ainsi, on peut dire que le jeu de données est très complet pour faire une étude de ces deux variables.

MRR Project 2018
Statistical Analysis and Description

- General nature of the problem

Link to data:
http://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set#
Our data include construction cost, sales prices, project variables, and economic variable corresponding to real estate single-family residential apartment in Tehran, Iran. The data includes 372 observations, and 109 variables.

Variable Group	Variable ID	Time lag number
Project date(Persian calender)	Start year Start quarter Completion Year Completion Quarter	N/A
Project physical and Financial variable	V-1 to V-10	N/A
Economic variable and indices	v-11 to V-29	1 to 5

Table 1: The summary of the variables in our data

As we look at the data and after loading the data in to R, we found that we have some problems.

1. The Label of the variable: We have redundancy of the name of variable in the Economic variable and Indices (They have the same notation in different Time lag Number).
2. The label of the Variable contain the symbol "-", and this causes problem in R, as they were confuse as the operator "-".
3. We have too many variables, therefore we need to find a good solution to drop the variables that are not useful in building our model.

- Explanatory Variable available

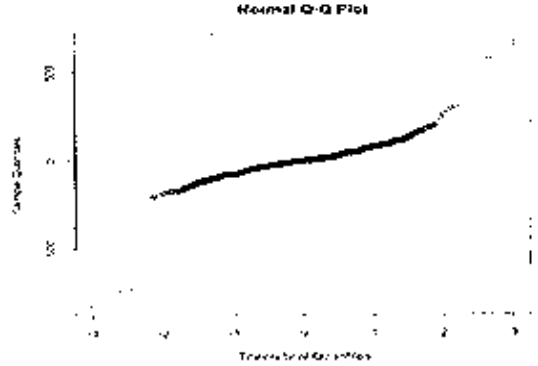


Illustration 1: The qq plot of V9 with the qq line

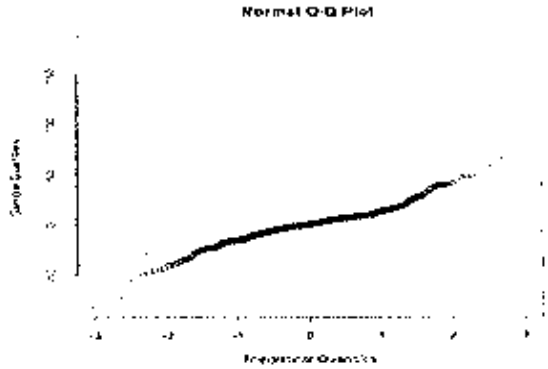


Illustration 2: The qq plot of V10, and the qq line

First, in order to be able to use the regular regression we need to make sure that our residual is gaussains distributed. We use the Shapiro.test, which gave us p-value<0.05. Then, we plot the qqplot and qqline to see the distribution of the data, and it shows a pretty good results indicated that our residuals are guassian, excluding some out-liners for both target variables.
To solve the problem of having too many explanatory variables, we use 3 regression methods to help choosing the explanatory variables. In order to do so, we compare 3 criterion as shown in the table below and find the compromise between the 3 criterion that we want to minimize.

	RSME	Number of variable	Error
Backward Regression	129.9474	32	135.9
Forward Regression	124.729	77	140.1
Stepwise Regression	128.2801	31	134

Table 2: Regular Regression for Target Variable V9. From this, we can see that the Stepwise regression give the best solution. Therefore, for V9, we keep 31 variables.

	RSME	Number of variable	Error
Backward Regression	20.60197	62	22.57
Forward Regression	20.43409	77	22.95
Stepwise Regression	20.60197	62	22.57

Table 3: Regular Regression with target variable V10. From this, we can know that the stepwise regression is give the best solution. Therefore, for V10, we keep 62 variables.

- Target variable and its link to explanatory variable

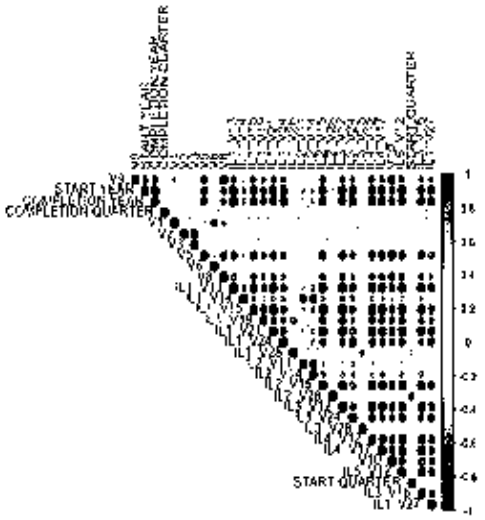


Illustration 3: This is a corrplot for V9 and the kept variables. We can see that V9 is strongly correlated to V8 and V5.

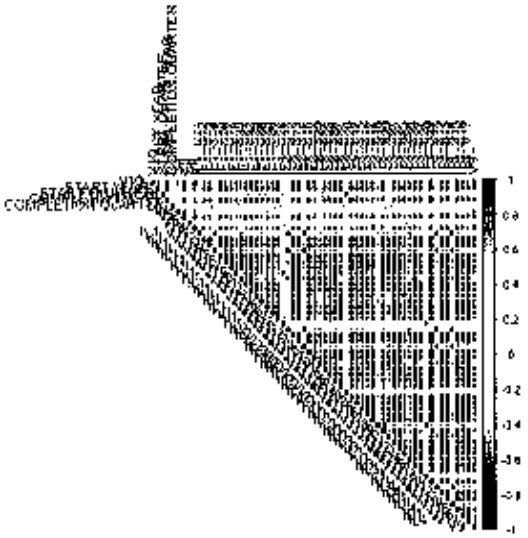


Illustration 4: This is a corrplot for V10 and the kept variables.

Our new variables for modeling of the target variable V9 are:
V9,STARTYEAR,COMPLETIONYEAR,COMPLETIONQUARTER,V1,V2,V3,V5,V6,V8,IL1_V13,IL1_V14,IL1_V15,IL1_V19,IL1_V21,IL1_V24,IL1_V28,IL2_V11,IL2_V14,IL2_V15,IL2_V28,IL3_V12,IL3_V24,IL3_V28,IL4_V12,IL4_V17,V10,IL5_V12,START QUARTER,IL3_V16,IL1_V2
Our new variables for modeling of the target variable V10 are:
V10,STARTYEAR,STARTQUARTER,COMPLETIONYEAR,COMPLETIONQUARTER,V4,V5,V6,V8,IL1_V11,IL1_V12,IL1_V13,IL1_V14,IL1_V15,IL1_V16,IL1_V17,IL1_V19,IL1_V20,IL1_V21,IL1_V22,IL1_V23,IL1_V24,IL1_V25,IL1_V26,IL1_V27,IL1_V28,IL2_V11,,IL2_V12,IL2_V13,IL2_V14,IL2_V15,IL2_V17,IL2_V20,IL2_V21,IL2_V25,IL2_V26,IL2_V27,IL2_V28,IL2_V29,IL3_V12,IL3_V15,IL3_V16,IL3_V17,IL3_V18,IL3_V19,IL3_V20,IL3_V21,IL3_V22,IL3_V23,IL3_24,IL3_V26,IL3_V27,IL3_V28,IL3_V29,IL4_V11,IL4_V14,IL4_V12,IL4_V13,IL4_V15,IL4_V16,IL4_V17,V9

projet_2pages_detail

My Dataset and the problem

This Data set includes construction cost, sale prices, project variables, and economic variables corresponding to real estate single-family residential apartments in Tehran, Iran.

It contains 8 project physical and financial variables(which is form v1 to v8 in data set), 19 economic variables and indices in 5 time lag numbers (from v11 to v29 in data set), and two output variables that are construction costs and sale prices.

Obviously, this dataset should be treated as an Linear Regression Problem, because the attributes and the goals are continuous, and we want to predict the construction costs and the the sale prices with the different physical & financial variables and the economic variables.

The dimension of the data: "observation" "variable"

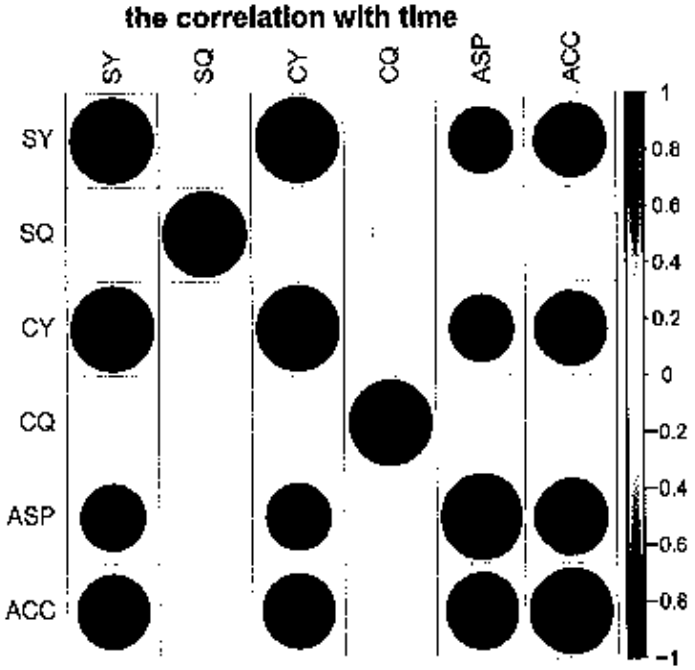
```
## [1] 372 109
```

The variables

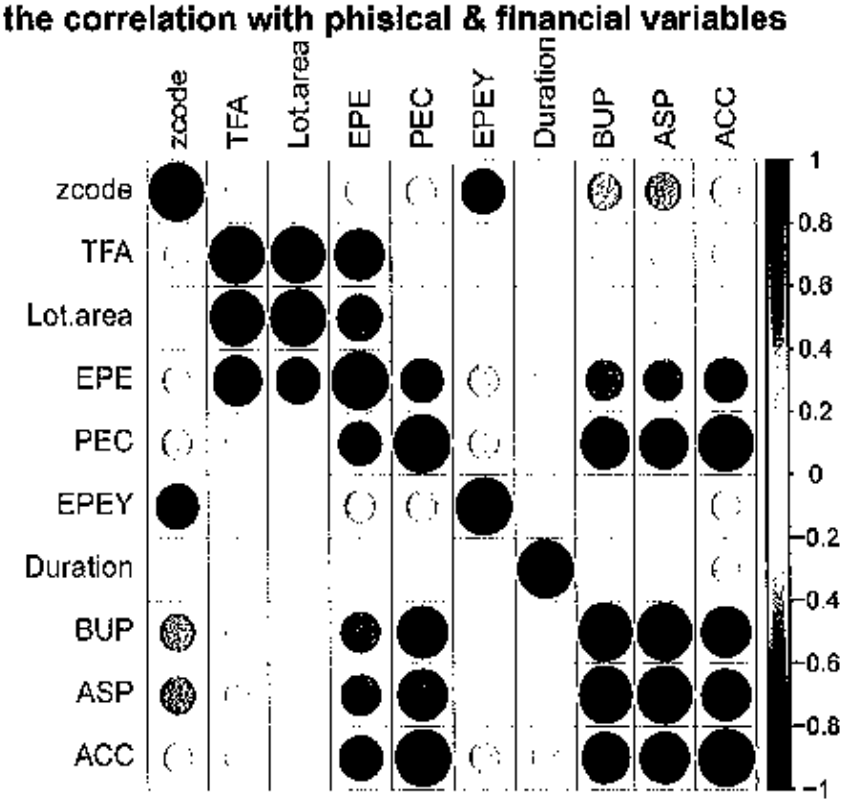
the target variables

The variables ASP and ACC are the target variables who is actual construction costs and sale prices. At first, we check the correlation bitween the target variables with others:

For the others variables, there are three groups:



So for this dataset, the first 4 variables are "start year", "start quarter", "completion year" and "completion quarter". These four variables describe when they build this building. And there is a relation between the variables about year and the targets. What's more, the interval between the start year and the end year maybe an important variable.



There are so many variables so I'd like to introduce the variales which has more correlation. We find that the variables zcode , EPE, PEC and BUP are more important, which are "Project locality defined in terms of zip codes", EPE: "Total preliminary estimated construction cost based on the prices at the beginning of the project", PEC :"Preliminary estimated construction cost based on the prices at the beginning of the project" and BUP:"Price of the unit at the beginning of the project per m2".

ECONOMIC VARIABLES AND INDICES

These variables are in 5 time lag numbers, so there are almost 95 variables. Each 19 variables describe the economic situation of the society for one instance. So perhaps they will influce the target variables, or ,which is more possible, there are many variable who are not interesting for the target variables.