

Predict real estate income in Rionegro's municipality



FINAL PROJECT SUBMITTED BY TEAM 53 FROM THE COURSE DATA SCIENCE FOR ALL, OFFERED BY MINTIC AND CORRELATION ONE

This document and the project developed as a web application was developed as a final project for the Data Science For All course.

The members of Team 53 are: Arbelaez David, Caballero Andrés, Gonzalez Sara, Rodriguez Oscar.

First release, September 2021

Index

1. Introduction
 - 1.1. Context
 - 1.2. Business Impact
 - 1.3. Methodology
2. Data
 - 2.1. Data Sources
 - 2.2. Exploratory Data Analysis
3. Model Design and selection
 - 3.1. K-means
 - 3.2. Multiple linear regression
 - 3.3. Selection of the best model
 - 3.4. Validation of statistical assumptions
 - 3.5. Final Model
 - 3.6. Forecast
4. Web application
 - 4.1. Mockup design
 - 4.2. Front-end
5. Conclusions and Future work

1. Introduction

1.1. Context

The principal income of the governments are the taxes and they are used to guarantee basic goods and services, and the main services are education, health and security. Also another use of them is to make the infrastructure better with elements such as streets, buildings, houses, electricity and sewerage. For that reason, the taxes are obligatory payments of the companies and persons. In Colombia there are a variety of taxes, all of them depending on the type of activity or consumption of products, the main ones are IVA (impuesto al valor agregado), business rent, GMF 4x1000 and others. Additionally, Colombia 's cities have other taxes such as ICA (impuesto de industria y comercio), predial and others.

The country has to make better use of this income, for that it has to create a budget. Colombia define this budget on the first three months of the year and this have to be send to the Treasury, on this place they determine if this budget continues and can be send to the Congress, specially to the economics area, then it is evaluate if there is a coherence between the objectives and the next incomes. But for the departments and cities in Colombia, it has to be an agreement between the municipal congressmen or department congressmen and the government.

All the time the expenses of a country change and most of the time it increases, for that reason there exists a need to create and change the tax reform. In Colombia, the entity in charge of these changes is the Republican Congressman and for the territory taxes is the department asamblea. Between 2010 and 2016, 4 laws have been created with the aim of acquiring more tax incomes and each of them have around 4 to 6 new items that create new taxes or decide to add more people to pay taxes.

One of the important taxes of the cities in Colombia is the real estate, this is the tax applied to the properties that are in the soil of the city. The Law 44 of 1990 determined the general regulatory guidelines of this tax and said the way of calculation is over the catastral value, technical value given from an expert that tries to approximate the real value of the properties, and define a specific rate depending on this value. This rate has to be between 1 to 16 per thousand. This is a progressive tax, that means the properties with major value have to pay more percentage.

Rionegro's Town (one of Antioquia's department municipalities) recognizes and suffers from the fact that sources of public resources are limited and the social needs tend to increase over time. Therefore, one of their aims is to predict incomes and then use this information to optimize the short and long term expenditure & investment planning and execution processes and thus solve the problem of planning and making investments without knowing what the actual income will be. An important source of this income is real estate. Project scope is limited to predict this specific income

With optimal prediction of income from real estate, among others, Rionegro's town will be able to plan and manage in advance their financial resources and direct and lead them in an efficient way to address their social investment and expenditure needs. The correct planning and management of the resources, investments and expenditure helps to meet and address population and community needs.

Literature review

- Agreement 023 from 2018 which establishes the fiscal regime for Rionegro's Town and rules the main taxes including, among others, the property tax.
- Guía ciudadana a la tributación y el gasto del Estado colombiano 2018.
- Tributación y pobreza en Colombia: un análisis desde la evolución del impuesto de renta y el índice de pobreza monetaria.

1.2. Business impact

Annually governments face important challenges in order to conduct an optimal income distribution over society, specifically in Colombia the annual budget is distributed throughout the ministries and something similar happens in the towns where the income is distributed not in ministries but in secretaries, regularly a town has 14 secretaries, each secretary has specific functions that aim to meet the different society needs. So, why is it important to accurately predict the income from taxes for governments?. Well, in simple terms if you know how much you will have, you will be able to structure a budget and plan how much you will spend. Let's keep in mind that about 49% of the income for governments comes from taxes, in that way we can say that we are trying to predict an important part of the total income.

Going deeper into the importance of predicting the income from taxes, we believe that efficient tax administration can highly improve the quality of life of the citizens and this statement has two approaches the first one comes from companies that will be expecting how taxes are invested aiming to get a city growth and in consequence a development of the economic environment, the other approach comes from citizens or society in general and this is probably the most important approach, beyond to answer the question, how the citizens are perceiving the efficiency from the

administration, we believe that a correct distribution of the income throughout social investments will increase the quality of life, so, how to make correct investments?, how to minimize risk in the cash flow of an investment?, how to assure that invoices will be paid?, how to plan how many projects start or how many investments we will be able to make?, how to assure the total development of a project and don't leave projects at the middle of the timeline? and finally how to estimate the social impact based on the investments made?. All the risks that imply answering these questions can be minimized throughout accurate forecasting models.

The American Economic Journal suggests the following: "Compliance with tax laws is important to keep the system working for all and supporting the programs and services that improve lives. One way to encourage compliance is to keep the rules as clear and simple as possible. Overly complicated tax systems are associated with high tax evasion. High tax compliance costs are associated with larger informal sectors, more corruption and less investment. Economies with simple, well-designed tax systems are able to boost businesses activity and, ultimately, investment and employment.", beyond trying to make easy rules for recollecting taxes, predict incomes for programs that save or improve lives could be one of the biggest impacts that we would like to achieve at the end of the paper.

Djankov, Simeon, Tim Ganser, Caralee McLiesh, Rita Ramalho and Andrei Shleifer. 2010. "The Effect of Corporate Taxes on Investment and Entrepreneurship." *American Economic Journal: Macroeconomics* 2 (3): 31–64.

With optimal prediction of income from property taxes, among others, Rionegro's town will be able to plan and manage in advance their financial resources and direct and lead them in an efficient way to address their social investment and expenditure needs. The correct planning and management of the resources, investments and expenditure helps to meet and address population and community needs.

1.3. Methodology

We will perform a univariate analysis through an exploratory data analysis, where we will find the measures of central tendency and the dispersion measures of the provided variables, in order to identify the behavior, trend and identify the completeness of the variables.

The univariate analysis will be extended to a bivariate one to examine the effects of causality or influence of each independent variable on the dependent variable of

interest (Rionegro property tax collection). The correlation and collinearity between the independent variables will also be observed, to rule out problems of this type that may affect the modeling that will be carried out later.

Finally, we will perform a multivariate analysis that could include a multifactorial technique or an analytical model that allows us to model and identify the behavior of Rionegro property tax collection through the known and historical data of the other variables, and then be able to predict their behavior in a future time period.

TIMELINE

Week	Deliverable	Details
Week 1	Team Creation	
Week 2	Project Assignment	Correlation One assigns the project to work in
Week 3	Project Description	-Understand the problem -Meet client and revisit problem -Be clear about what we want to solve
Week 4	Project Scoping Definition	-Share concerns with client on scope and Datasets requirements
Week 5	Final Datasets agreed and provided	-Define datasets based on concerns and business needs
Week 6	Basic EDA, Cleaning of datasets completed	
Week 7	In-depth EDA, jupyter analysis, mockup of frontend	
Week 8	First application approach and definition of model to use	
Week 9	Custom predicting	-Try with time series

	model implemented	models, compare and evaluate them
Week 10	Front End design	
Week 11	Application infrastructure complete	
Week 12	Datafolio completed	Includes: Application, final document, video and presentation

2. Data

2.1. Data Sources

Our client has initially provided two datasets consolidating property tax invoices and payments. Their feature details are listed below:

CAMPO	TIPO	DESCRIPCION
cc_tarifa	INTEGER	Milaje o valor que aplicado a la base gravable determina el monto del impuesto
cc_estrato_se	INTEGER	Numero asignado por catastro que se utiliza para determinar la tarifa
barrio	INTEGER	Lugar donde está situado el predio
area_terr	STRING	Numero de metros cuadrados que mide el terreno
area_terr_comun	STRING	Numero de metros cuadrados que se comparte en una propiedad horizontal
area_const	STRING	Número de metros cuadrados de la construcción
area_const_comun	STRING	Numero de metros cuadrados construidos que se comparten en una propiedad horizontal
vlr_terr	STRING	Valuación del inmueble sin construcción
vlr_terr_comun	STRING	Valuación del inmueble sin construcción que se comparte en una propiedad horizontal
vlr_const	STRING	Valuación de la construcción
vlr_const_comun	STRING	Valuación de la construcción que se comparte en una propiedad horizontal
vlr_tot_avaluo	INTEGER	Es la suma del valor del terreno y la construcción incluyendo áreas comunes
nro_doc_id_catastral	INTEGER	Documento de identificación del inmueble
propietario	FLOAT	Nombre de la empresa o del inmueble
avaluó por proindiviso	STRING	Porcentaje del inmueble que le corresponde a cada propietario
impuesto	STRING	Valor a pagar por impuesto predial
% ambiental	STRING	Porcentaje de impuesto predial correspondiente a cornare
% bomberil	STRING	Porcentaje de impuesto predial correspondiente a bomberos
total	STRING	Valor total a pagar (suma de impuesto más porcentajes)

As it may be noted in the above description of data set attributes, the data sets provide information for existing properties or units that are currently subject to the property tax. The client only provided this information for the last 5 years. We believe that we need the same for more historical data.

The information provided may suffice to predict what existing properties or units will pay in the future, however, we need to consider also that a considerable increase of property tax income comes after new properties are taxed as result of property splits and new apartments, houses and commercial units are built. Let's call all these "New Units". Creation of New Units usually depend other business and context variables which can be categorized in the following groups:

- Demographic behaviour and population
- Land Offer and availability
- Occupancy of Existing Units
- Economic growth
- GIS location of existing Units

As the effect of new units in property tax income may be significant, we also need historical data sets that include information and attributes related to the above set of variables in such a way that they are incorporated into the analysis and model.

External data sources

Some external data sources that may add business context and additional data variables to the problem and solution include:

- Economic growth:
 - PIB
 - IPC
 - Unemployment rate
 - IPVU
 - TRM

Having these variables the team decided to construct a new database, where the total value of the bills are grouped by year and by this way try to identify the trend of the main feature. The variables on this database are: year, total real estate, total real estate divided, IPVU (índice de precios de la vivienda usada) real total, IPVU real medellin, IPVU nominal total, IPVU real VIS, IPVU nominal VIS, PIB (Producto Interno Bruto), PIB variación total anual, PIB ph variación anual, inflación, tasa de ocupación, tasa de desempleo, tasa de ocupación 7am, tasa desempleo 7a, DTF , tasa de intervención banco de la república, trm, tasa interés real, variación anual pib.

2.2. Exploratory Data Analysis

As an introduction to the Exploratory Data Analysis we will have a journey that will start by exposing the most important insights found throughout the exercise in order to give to the reader a brief summary of the data provided by the client, then we will go deeper into the analysis of specific variables and finally we will conclude about which variables worth to include into future models taking into account how complete variables are and how are they initially related to the response variable.

Making an initial Exploratory Data Analysis we have find the following insights that help to describe the data available:

- We assume as a unique identifier the 'Property Consecutive'.
- We have information on 69.308 properties in at least one of its years.
- From 91.28% of the properties we have information from the last 5 years (Table 1).
- Of the 8.72% of the properties we can assume that they appeared as new property units as a result of some transformation (eg: A lot can go from having one owner to two).
- 25.33% of the properties have had more than one owner in the last 5 years (Table 2).
- It is of our interest only the information of the property not of the owners of the property, therefore we want to unify the owners of the same property as a single property.
- The number of properties that entered the database in the last five years was an average of 1,516 per year. In 2017 they entered compared to the previous year 1965, in 2018 compared to 2017 they entered 1,274, in 2019 compared to 2018 1721 entered, in 2020 compared to 2019 1,721 entered.
- The number of properties that came out of the database in the last five years was an average of 52 per year. In 2017, 121 came out compared to the previous year, in 2018 compared to 2017 there were 0 (to be reviewed), in 2019 With respect to 2018, 18 came out, in 2020 with respect to 2019, 68 came out.

# of years with data	% of properties with data
1	1,56%
2	2,43%
3	1,90%
4	2,83%
5	91,28%
Total	100,00%

Table 1

	Percentage
Uniquer owner	74,67%
More than one owner	25,33%
Total	100,00%

Table 2

The feature `cc_tarifa` refers to the percentage of thousands that is applied in the tax value. This tax has been determined on the estatuto tributario acuerdo 023 de 2018. On the table 3 one important idea to be noticed is there is a change in 2017 because it is seen as NaN value after and before that year. Also the institute has a strategy for the unknown value they decided to assign it as 999 and knowing that it can be seen a majority of missing values and this value is not confiable for us (Table 3).

	cc_tarifa2016	cc_tarifa2017	cc_tarifa2018	cc_tarifa2019	cc_tarifa2020
2	NaN	26134.0	33150.0	34516.0	35529.0
5	9899.0	11745.0	29415.0	30229.0	30304.0
6	308.0	NaN	NaN	NaN	NaN
7	1483.0	NaN	NaN	NaN	NaN
8	5253.0	23157.0	1573.0	1590.0	1429.0
9	10668.0	NaN	NaN	NaN	NaN
10	23097.0	NaN	NaN	NaN	NaN
11	1935.0	2183.0	789.0	767.0	711.0
12	15175.0	NaN	NaN	NaN	NaN
13	2227.0	NaN	NaN	NaN	NaN
14	10811.0	24650.0	24662.0	24691.0	24502.0
15	1008.0	NaN	NaN	NaN	NaN
16	4515.0	21.0	21.0	NaN	NaN
33	2924.0	2887.0	2842.0	2973.0	3034.0
999	4017.0	4494.0	4447.0	4373.0	4739.0

Table 3

Another categoric variable is the Estrato which is the socioeconomic segmentation, this one has values between 1 to 6 defined for the government. On the table 4, one important aspect to notice is that there is an estrato 0; it is an error also it is not confiabile for the analysis of this value. Additionaly, in this case for the municipio of Rio Negro it is seen that the majority of the properties are situated in estrato 3 or 4 (Table 4).

	Estrato_2016	Estrato_2017	Estrato_2018	Estrato_2019	Estrato_2020
Estrato					
0	35566	35760	35845	36239	36224
1	1419	1577	1689	1838	1862
2	6438	6659	6869	7073	7101
3	24057	24409	24599	25207	25925
4	16850	17634	18502	19191	19560
5	3410	3726	3883	3975	4010
6	1834	1730	1735	1863	1788

Table 4

In the following chart it is seen that the total number of predios has been increasing over the time and for 2020 it finished at 69.240 predios that paid for their tax and it can be seen that it increases about 2000 each year (Table 5).

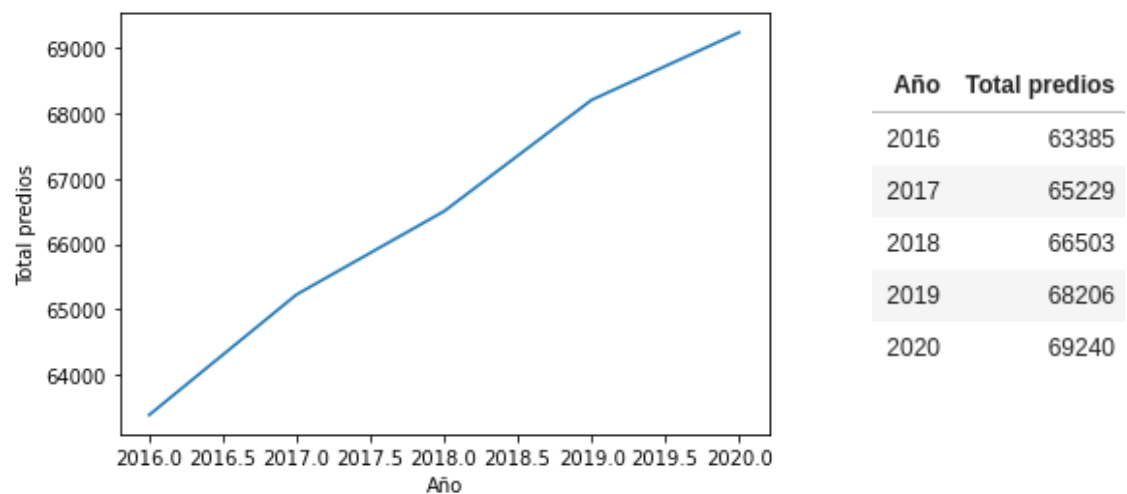


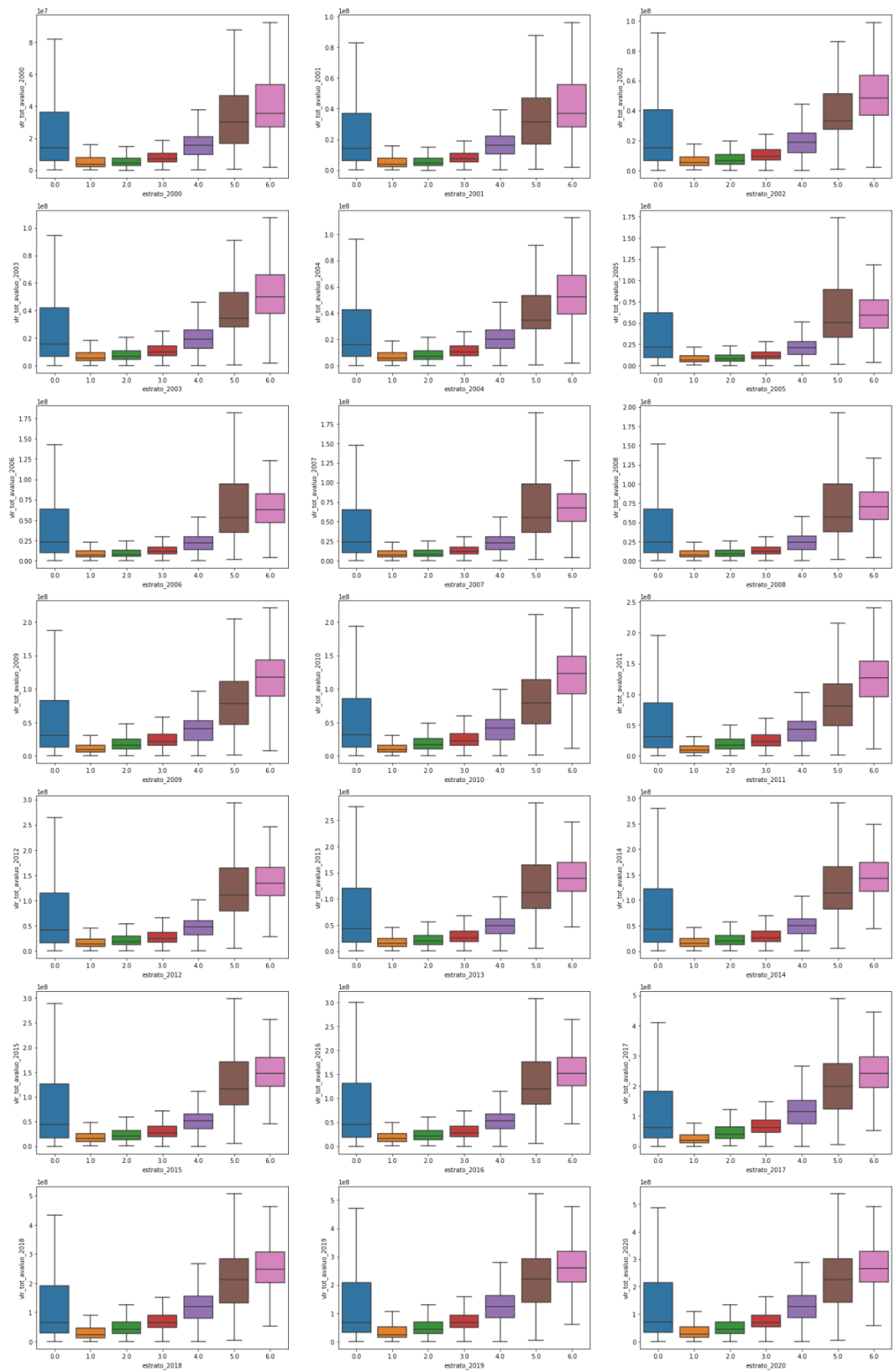
Table 5

Another important feature of the institute is the `vlr_const` that is the total value of the constructed area. On the next chart it is seen that all of the estratos have a minimum of 0, so it is not common to have a construction with this value that causes the project to drop that value or ask for better information about that. Another insight of the 2016 chart is that for estrato 1 that is the first socioeconomic category that has the minimum values. Also it can be seen that the median on all of the estratos is different and for estrato 5 and 6 have the maximum median.

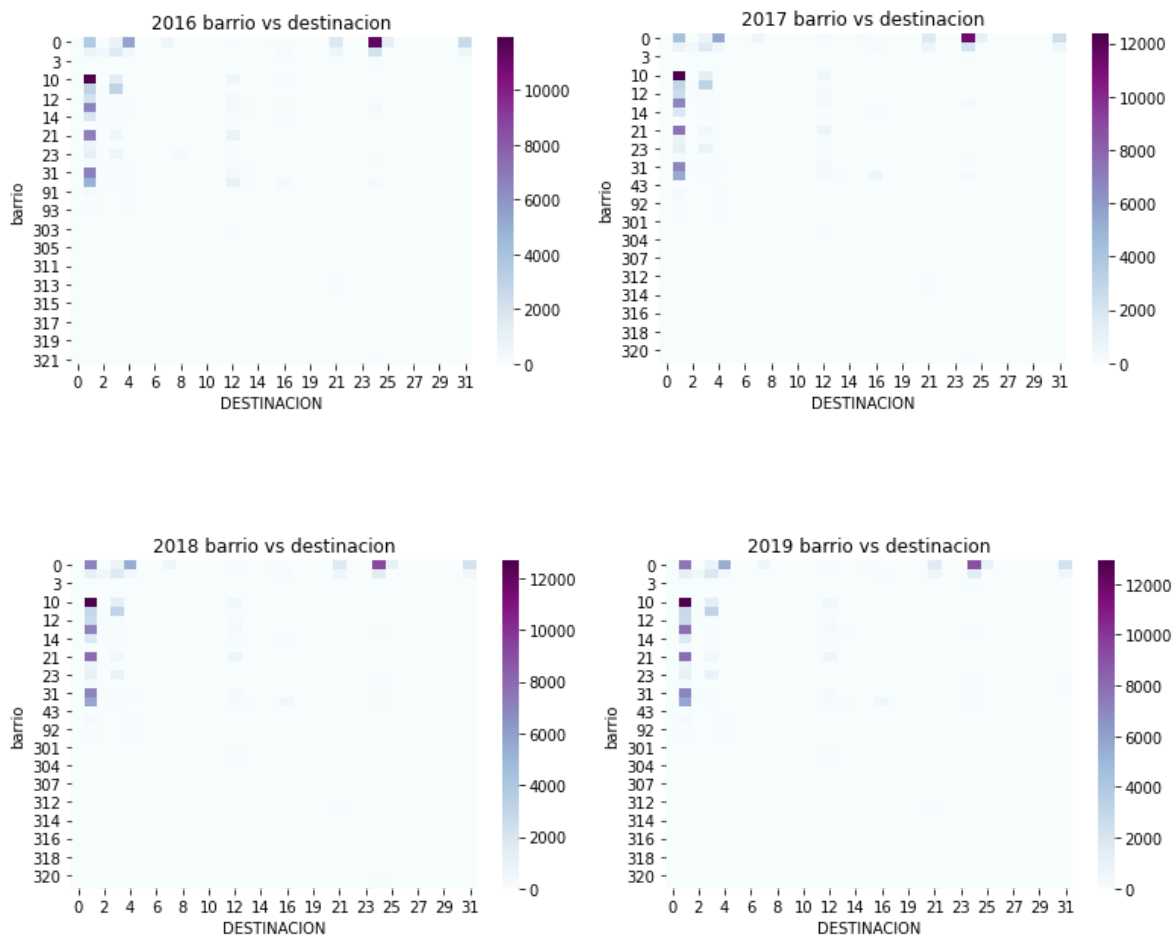
Comparing 2016 and 2017 it can be seen that the biggest change has been estrato 4 that has increased about 0.25 their median. Showing for 2017 the estrato 4 could have a rise in the area value or increase in the amount of area constructed.

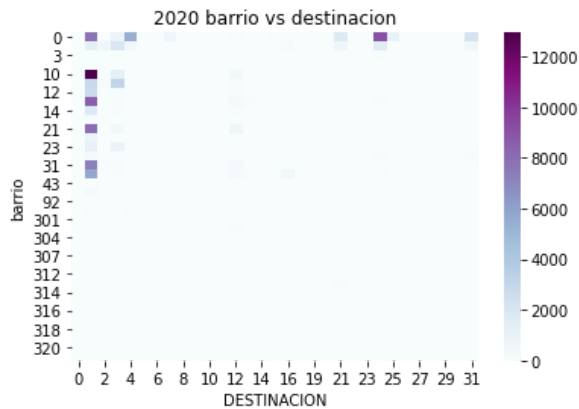
Comparing all the following charts it can be seen that Estrato 6 has increased their maximum value in each year. Also, the estratos with more constructed value are 5 and 6, and this is a normal representation of the estratos of Colombia, where these estratos are the most valued.

Total constructed value by Estrato.



The next charts show how it is distributed the economic destination for each barrio, in other words it is the number of properties for each destination on the barrios, where the darker the color the more properties are in there. In 2016 it is seen that the destinacion (economic destination of the place) with higher participation on the barrios is the 25 that is the cattle raising, one important thing to have present is that Rio Negro is known for the amount of area that is for raising cows. Another important insight is that destinacion 2 is present in barrios 0,10,12,14,21 and 31 and this could be seen as the industrial areas because the destinacion 2 is for the industries.

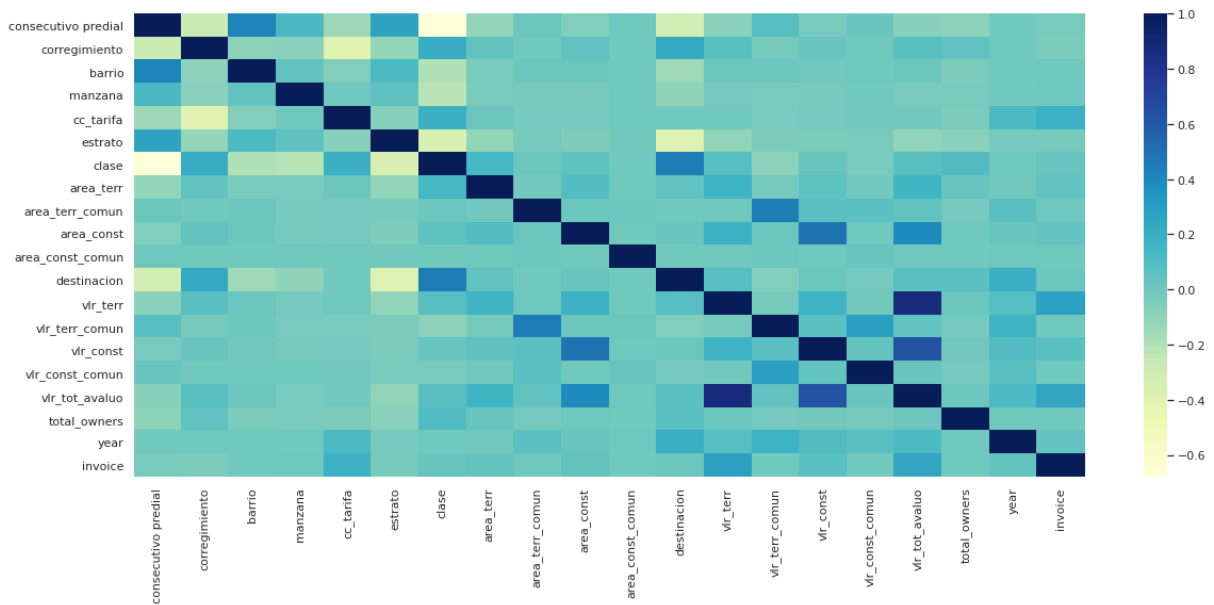




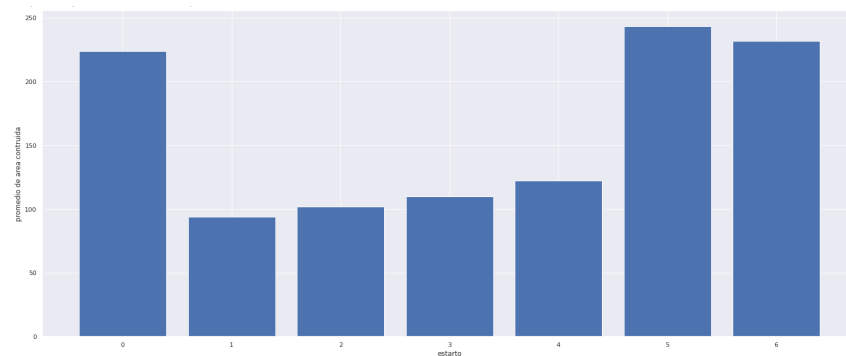
Over the time it is seen that the major participation of the destination is the cattle raising, in each year this one has the biggest participation. Also it can be seen is that the barrio 10 is the principal industry area of Rio Negro.

In this moment of the project we have to say that our data is separated by years, the main idea of the team is to make only one dataframe where we can have the changes over the time of each estate. For that we need to consolidate and clean each year of information to do our feature engineering.

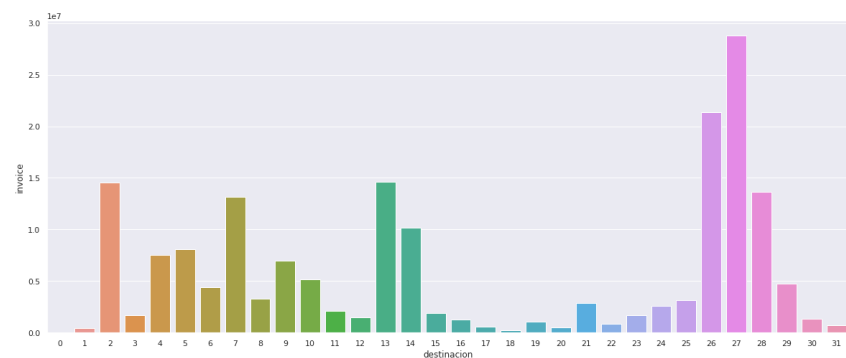
On of the most important charts in a EDA could be a heatmap showing the correlation among variables as wee see in the following chart there an important correlation between the variable "invoice" which is the value to be charged to the inhabitant and group conformed by: cc_tarifa and vlr_total_avaluo this make totally sense and is consistent with business context, another expected correlation and that we confirm with the heatmap is the correlation between area_const and valor_tot_avaluo, the larger your property is the more expensive it will be.



Going through the analysis of the estratos we have seen that the largest properties are located in the highest estratos as we see in the following figure. Also, we see that large properties are in the estrato 0, it's probably because there are many uninhabited lots.



It is interesting to see that destinacion 27 which are properties that are dedicated to build educational centers pay on average the highest taxes.

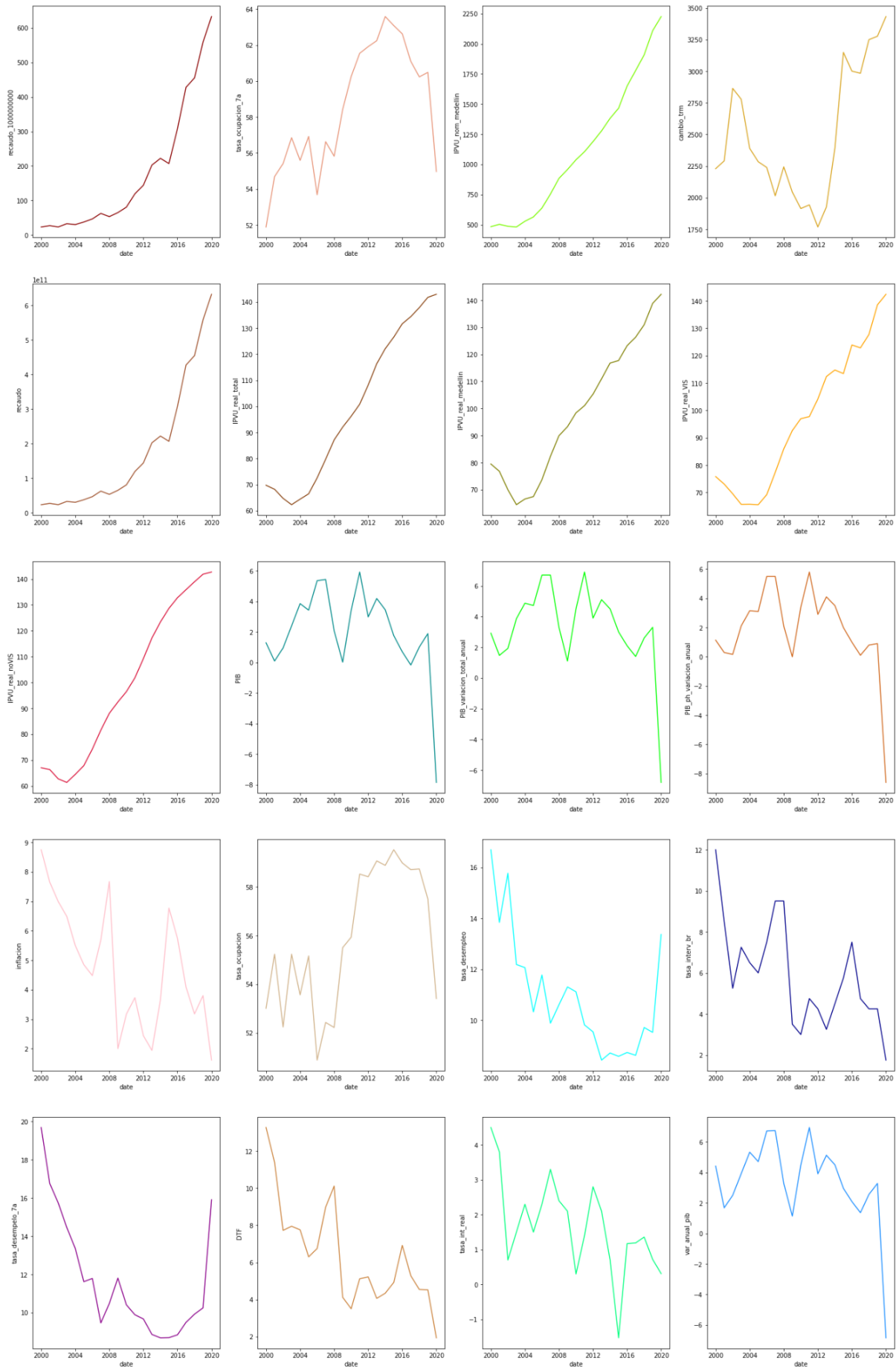


On the next imagen, it can be seen the behaviour of the features over the time, this database has the variables used to construct the final model. As the main purpose of this model is to predict the next income of the real estate tax so it is important to understand this behaviour.

The dependent variable is the real estate and it is shown that it has a exponential behaviour, starting in 2000 at 23.56 thousand million and finishing around 632.43 thousand millions. Additionally, looking the other variables the one that have similar behaviour is Medellin's IPVU, which starts at 483.88 and finishes approximately at 2225.75.

Another important variable for understanding how the country has improved is the PIB, because if the country is having a good moment it can be interpreted as a good moment for all the taxes on the country. It can be seen that after 2000 it started to increase until 2007 around to 5.43, then it slightly decreases on 2008 due to a world economic crisis and for 2010 it increases to 2.07 but on for the last years it extremely decreases, finishing with a negative value of 7.84 and it can be explained for the pandemic around the world.

Finally, the unemployment rate can give an idea how much people are working and this can be correlated to the amount of taxes paid. But, for this variable on the first view there is no correlation between them, because this variable decreases rapidly until 2017 where it finished at 8.63 and then it dramatically increase until 13.37 in 2020.



3. Model Design and selection

3.1. K-means clustering algorithm

Once the context and the business problem has been exposed and EDA has been performed, we introduce the concept of the K-means algorithm. K-means clustering is an unsupervised machine learning model that aims to make inferences from datasets in order to make categories or groups with representative characteristics in each one, as it is an unsupervised algorithm we don't expect a "correct" answer.

AndreyBu suggests that the objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset. A cluster refers to a collection of data points aggregated together because of certain similarities.

The algorithm defines k numbers of centroids, and then locates every data point to its nearests clustering, always trying to make the distances inside the cluster as small as possible.

Dr. Michael J. Garbade describe how the algorithm works:

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

We consider that trying to explore a clustering algorithm will give to the project an extra level of accuracy taking into account that a specific forecast model could be developed for each group of properties.

It's important to note that for future releases it will be essential to join the properties dataset with the invoices dataset.

The above point was not able to be covered in the first release.

Going ahead with the results we find the following:

Using the elbow method an optimal number of clusters is 4 and these are the most important characteristics for each cluster:

Cluster 0:

- 3.05% of the properties belong this cluster
- **total_owners** is in average 403% greater : mean of 8.47 against 1.68 globally
- 44.18% of the cluster has 4 for **destinacion** (against 19.61 % globally)
- 39.13% of the cluster has 0 for **barrio** (against 19.40 % globally)

Cluster 1:

- 62.77% of the properties belong this cluster
- 97.70% of the cluster has 1 for **destinacion** (against 63.90 % globally)
- 64.68% of the cluster has 3.0 for **estrato** (against 43.66 % globally)
- 34.33% of the cluster has 10 for **barrio** (against 23.24 % globally)

Cluster 2:

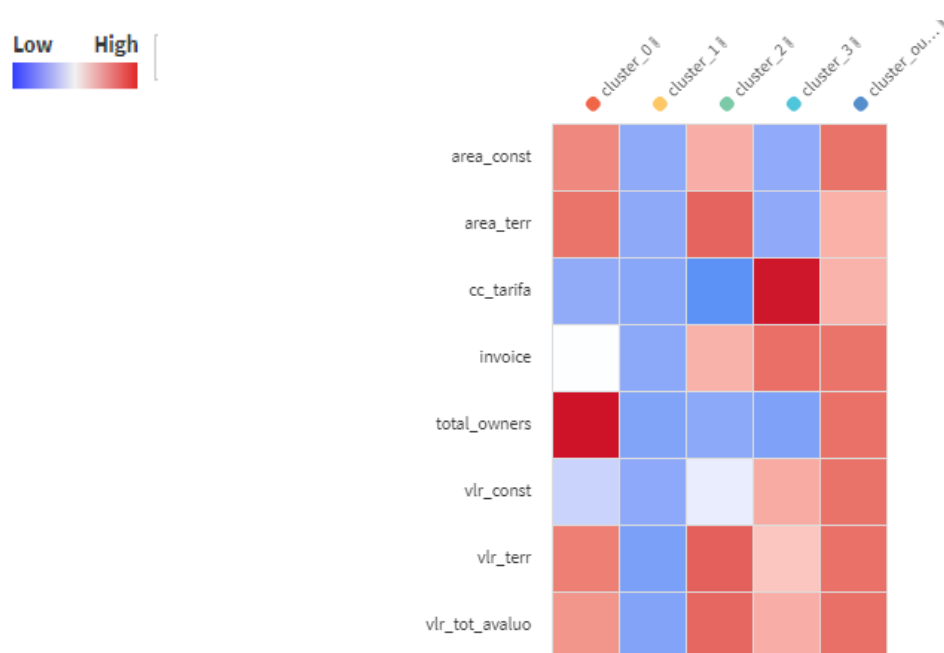
- 17.74% of the properties belong this cluster
- 98.20% of the cluster has 4 for **destinacion** (against 19.61 % globally)
- 78.50% of the cluster has 0 for **barrio** (against 19.40 % globally)
- 69.01% of the cluster has 0.0 for **estrato** (against 28.58 % globally)

Cluster 3:

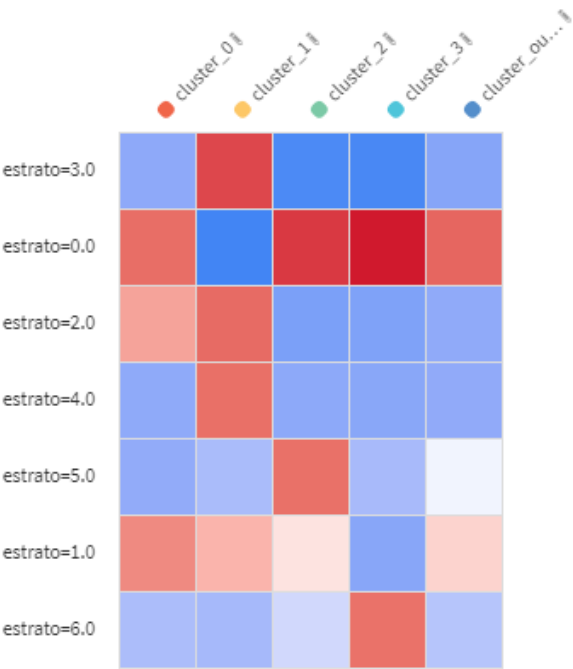
- 15.64% of the properties belong this cluster
- 49.84% of the cluster has 3 for **destinacion** (against 8.11 % globally)
- 46.44% of the cluster has 11 for **barrio** (against 13.43 % globally)
- 82.38% of the cluster has 0.0 for **estrato** (against 28.58 % globally)

Only 0.79% of the properties were considered outliers

The following heatmap helps us to visualize what are the most relevant numeric features for each cluster.

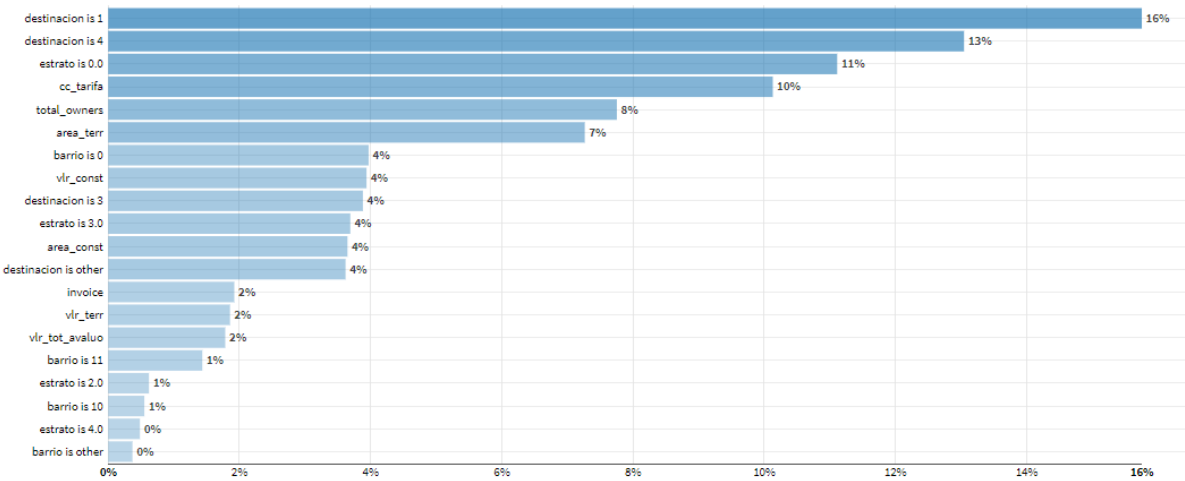


One important feature for the business is the stratum here is how representative each stratum is across the clusters.



Finally we expose what are the most relevant variables when it comes to clustering the properties.

In the chart below we can see that properties could be clustered mainly taking into account its destination and stratum which make us conclude that first the result make sense taking into consideration the business context and second the most relevant features highlighted at the beginning of the results for each cluster are consistent because they include destination and stratum



3.2. Multiple linear regression

The multiple linear regression problem is considered when the study variable depends on more than one explanatory or independent variable. This model generalized linear regression allowing the mean function $E(y)$ (Mathematical expectation of variable y) to depend on more than one explanatory variable.

The linear model:

Let y denotes the dependent (or study) variable that is linearly related to k independent (or explanatory) variables X_1, X_2, \dots, X_k through the parameters $\beta_1, \beta_2, \dots, \beta_k$ and we write

$$y = X_1\beta_1 + X_2\beta_2 + \dots + X_k\beta_k + \varepsilon$$

This is called the multiple linear regression model. The parameters $\beta_1, \beta_2, \dots, \beta_k$ are the regression coefficients associated with X_1, X_2, \dots, X_k respectively and ε is the random error component that corresponds to the difference between the fitted linear regression and the observed data. There may be several reasons for such a difference, among which is the joint effect of variables that are not included in the model and could also affect the behavior of the variable y , the random factors that can not be taken into account in the model, etc.

The j^{th} regression coefficient β_j represents the expected change in y per unit change in the j^{th} independent variable X_j . Assuming $E(y)$ and $\beta_j = \frac{\partial E(y)}{\partial X_j}$.

A model is said to be linear when it is linear in parameters. In such a case $\frac{\partial y}{\partial \beta_j}$ (or equivalently $\frac{\partial E(y)}{\partial \beta_j}$) should not depend on any β 's. For example:

- i. $y = \beta_0 + \beta_1 X$ is a linear model as it is linear in the parameters.
- ii. $y = \beta_0 X^{\beta_1}$ can be written as

$$\begin{aligned}\log \log (y) &= \log(\beta_0) + \beta_1 \log(X) \\ y^* &= \beta_0^* + \beta_1 x^*\end{aligned}$$

which is linear in the parameter β_0^* and β_1 , but nonlinear in variables $y^* = \log \log (y)$, $x^* = \log(X)$. So it is a linear model.

- iii. $y = \beta_0 + \beta_1 X^{\beta_2}$ is nonlinear in the parameters and variables both. So it is a nonlinear model.

Prediction model for income from real estate taxes in the municipality of Rionegro

Se pretende realizar un modelo de regresión lineal múltiple que estime el recaudo anual de los ingresos por impuesto predial en el municipio de Rionegro, con base en indicadores macroeconómicos que puedan explicar el comportamiento de esta variable.

Y : Recaudo anual de los ingresos por impuesto predial en el municipio de Rionegro dividido en 1.000.000.000 (Recaudo).

X_1 : (IPVU_real_total)

X_2 : (IPVU_real_medellin)

X_3 : (IPVU_nom_total)

X_4 : (IPVU_nom_medellin)

X_5 : (IPVU_real_VIS)

X_6 : (IPVU_real_noVIS)

X_7 : (IPVU_nom_VIS)

X_8 : (IPVU_nom_noVIS)

X_9 : (PIB)

X_{10} : (PIB_variacion_total_anual)

X_{11} : (PIB_ph_variacion_anual)

X_{12} : (Inflacion)

X_{13} : (Tasa_ocupacion)

X_{14} : (tasa_desempleo)

X_{15} : (tasa_ocupacion_7a)

X_{16} : (tasa_desempelo_7a)

X_{17} : (DTF)

X_{18} : (tasa_interv_br)

X_{19} : (cambio_trm)

X_{20} : (tasa_int_real)

X_{21} : (var_anual_pib)

Las variables X_k corresponden a las variables independientes del modelo, es decir, por medio de estas se describirá el comportamiento del *Recaudo*. Las variables que resulten ser significativas para la estimación del modelo, serán explicadas más adelante.

3.3. Selection of the best model

- A. Se tuvo en cuenta que las variables X_1 a X_8 presentan una correlación muy alta entre ellas por ser indicadores basados en el IPVU, por lo cual se seleccionó la variable que mayor correlación tiene con la variable *Recaudo*:

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Y	1	0,899	0,914	0,947	0,959	0,919	0,891	0,957	0,943
X_1	0,899	1	0,992	0,989	0,981	0,99	0,999	0,981	0,989
X_2	0,914	0,992	1	0,983	0,984	0,997	0,988	0,983	0,982
X_3	0,947	0,989	0,983	1	0,997	0,984	0,987	0,997	0,999

X_4	0,959	0,981	0,984	0,997	1	0,984	0,979	0,999	0,996
X_5	0,919	0,99	0,997	0,984	0,984	1	0,985	0,987	0,982
X_6	0,891	0,999	0,988	0,987	0,979	0,985	1	0,978	0,986
X_7	0,957	0,981	0,983	0,997	0,999	0,987	0,978	1	0,996
X_8	0,943	0,989	0,982	0,999	0,996	0,982	0,986	0,996	1

Según la tabla anterior la variable X_4 (*IPVU_nom_medellin*) es la que presenta mayor correlación lineal con la variable Y (*Recaudo*), por lo que se dejará únicamente esta variable dentro de todas las variables basados en el indicador IPVU evitando que posteriormente el modelo presente problemas de correlación entre las variables independientes.

- B. Se tuvo en cuenta que las variables X_9 , X_{10} , X_{11} y X_{21} presentan una correlación muy alta entre ellas por ser indicadores basados en el PIB, por lo cual se seleccionó la variable que mayor correlación tiene con la variable *Recaudo*:

	Y	X_9	X_{10}	X_{11}	X_{21}
Y	1	-0,583	-0,575	-0,583	-0,613
X_9	-0,583	1	0,997	0,994	0,987
X_{10}	-0,575	0,997	1	0,993	0,993
X_{11}	-0,583	0,994	0,993	1	0,98
X_{21}	-0,613	0,987	0,993	0,98	1

Según la tabla anterior la variable X_{21} (*var_anual_pib*) es la que presenta mayor correlación lineal inversa con la variable Y (*Recaudo*), por lo que se dejará únicamente esta variable dentro de todas las variables basados en el indicador PIB evitando que posteriormente el modelo presente problemas de correlación entre las variables independientes.

- C. Se tuvo en cuenta que las variables X_{13} , X_{14} , X_{15} y X_{16} presentan una correlación muy alta entre ellas por ser indicadores de ocupación y desempleo, por lo cual solo una de estas podría ser tenida en cuenta en el modelo:

	Y	X_{13}	X_{14}	X_{15}	X_{16}
Y	1	0,436	-0,349	0,375	-0,247

X_{13}	0,436	1	-0,707	0,914	0,608
X_{14}	-0,349	-0,707	1	-0,841	0,958
X_{15}	0,375	0,914	-0,841	1	-0,826
X_{16}	-0,247	0,608	0,958	-0,826	1

Según la tabla anterior la variable X_{13} (*tasa_ocupacion*) y X_{15} (*tasa_ocupacion_7a*) presentan una correlación muy alta, al igual que las variables X_{14} (*tasa_desempleo*) y X_{16} (*tasa_desempleo_7a*), por lo cual en principio las variables opcionales para quedar en el modelo serían X_{13} (*tasa_ocupacion*) y X_{14} (*tasa_desempleo*) las cuales presentan mayor correlación lineal con la variable Y (*Recaudo*), sin embargo, estas dos variables presentan una correlación de -0,7, razón por la cual se decide dejar únicamente la variable X_{13} (*tasa_ocupacion*).

- D. Se iniciará con un modelo de regresión lineal múltiple que incluya todas las variables mencionadas a continuación y se irán descartando paso a paso variables por medio del método **Backward Stepwise Regression**, es decir, se descartaron las variables menos influyentes bajo el criterio del mayor p-valor.

$$Y = \beta_4 X_4 + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} + \varepsilon$$

- i. Se inicia con un modelo de 8 variables, obteniendo los siguientes resultados:

Variable	Estimate	t -value	p-value
IPVU_real_total	2,94E-01	11.508	3,45E-08
inflacion	-5,75E+04	-0,411	0,687476
tasa_ocupacion	-8,11E+03	-6,936	1,03E-05
DTF	6,05E+06	0,431	0,673613
tasa_interv_br	-1,59E+06	-1.593	0,135077
cambio_trm	1,25E+02	4,907	0,000287
tasa_int_real	-2,53E+06	-0,17	0,867512
var_anual_pib	4,36E+05	1,146	0,272484

De acuerdo con la tabla anterior, la variable *tasa_int_real* se elimina del modelo.

- ii. Después de eliminar la variable X_{20} (*tasa_int_real*) obtenemos los siguientes resultados:

Variable	Estimate	t -value	p-value
IPVU_real_total	2,93E+02	11,936	1,00E-08
inflacion	-3,38E+04	-3,823	0,00186
tasa_ocupacion	-8,18E+03	-7,701	2,13E-06
DTF	3,67E+06	3,830	0,00184
tasa_interv_br	-1,64E+06	-1,767	0,09901
cambio_trm	1,27E+02	5,515	7,61E-05
var_anual_pib	4,48E+05	1,244	0,23382

De acuerdo con la tabla anterior, la variable *var_anual_pib* se elimina del modelo.

- iii. Después de eliminar la variable X_{21} (*var_anual_pib*) obtenemos los siguientes resultados:

Variable	Estimate	t -value	p-value
IPVU_real_total	2,84E+02	11,952	4,56E-09
inflacion	-3,22E+04	-3,615	0,00255
tasa_ocupacion	-7,21E+03	-9,822	6,32E-08
DTF	3,09E+06	3,625	0,0025
tasa_interv_br	-1,09E+06	-1,311	0,20967
cambio_trm	1,14E+02	5,430	6,96E-05

De acuerdo con la tabla anterior, la variable *tasa_interv_br* se elimina del modelo.

- iv. Después de eliminar la variable X_{18} (*tasa_interv_br*) obtenemos los siguientes resultados:

Variable	Estimate	t -value	p-value
IPVU_nom_medellin	0,27384	11,905	2,31E-09
inflacion	-35,587	-4,090	0,000855
tasa_ocupacion	-7,217	-9,624	4,67E-08
DTF	2268,63	3,836	0,001458
cambio_trm	0,1208	5,802	9,00E+00

En este paso, tenemos se obtuvo un modelo donde todas las variables son significativas al 5% de significancia, con un R² Ajustado del 98,67%, AIC de 208,3246, sin embargo, al ser un modelo con 5 variables, se intentará buscar un modelo más parsimonioso.

- E. Partiendo del modelo anterior, se revisarán las correlaciones entre las variables del modelo, validando que no existan correlaciones entre estas.

Variable	IPVU_nom_medellin	inflacion	tasa_ocupacion	DTF	cambio_trm
IPVU_nom_medellin	1	-0,567	0,581	-0,682	0,608
inflacion	-0,567	1	-0,396	0,882	0,055
tasa_ocupacion	0,581	-0,396	1	-0,506	0,163
DTF	-0,682	0,882	-0,506	1	-0,234
cambio_trm	0,608	0,055	0,163	-0,234	1

Según la tabla anterior, las variables *IPVU_real_total* y *DTF* presentan una correlación por encima de 65%, y las variables *inflación* y *DTF* presentan una correlación de 88%. Por lo anterior se eliminará las variables *DTF* del modelo, obteniendo los siguientes resultados:

Variable	Estimate	t -value	p-value
IPVU_nom_medellin	0,267260	8,668000	0,000000
inflacion	-8,031310	-1,217000	0,240310
tasa_ocupacion	-5,628990	-6,698000	0,000004
cambio_trm	0,093650	3,558000	0,002420

Según la tabla anterior, la variable *inflación* pierde significancia al quitar la variable *DTF*, por lo cual se quitará del modelo, obteniendo los siguientes resultados:

Variable	Estimate	t -value	p-value
IPVU_nom_medellin	0,296530	15,174000	0,000000
tasa_ocupacion	-5,974660	-7,455000	0,000001
cambio_trm	0,072950	3,584000	0,002120

Según la tabla anterior se tiene un modelo con 3 variables estadísticamente significativas al 5%, con un R²-Ajustado de 97,53%, un BIC de 223,95 y un AIC 219,77.

- F. Teniendo en cuenta que dentro del modelo anterior se encuentra la variable *tasa_ocupación*, se probará el mismo modelo cambiando esta variable por *tasa_ocupacion_7a*, ya que a pesar de ser descartada en el paso 1.3 debido a

presentar un poco menos de correlación con la variable *Recaudo*, es una variable construida de forma muy similar a *tasa_ocupacion* y que podría brindar un aporte significativo al modelo. Al realizar el cambio se obtuvieron los siguientes resultados:

Variable	Estimate	t -value	p-value
IPVU_nom_medellin	0,30228	17,277	0,000000
tasa_ocupacion_7a	-5,68724	-8,562	0,000000
cambio_trm	0,06976	3,983	0,000872

Según la tabla anterior se tiene un modelo con 3 variables estadísticamente significativas al 5%, con un R2-Ajustado de 98,29%, un BIC de 219,42 y un AIC 215,24.

De acuerdo con lo anterior, el modelo mencionado en el punto 1.6 tiene un punto más de ajuste en el indicador R2-Ajustado y tiene menor AIC y BIC, por lo cual será seleccionado como mejor modelo.

3.4. Validation of statistical assumptions

El modelo presentado en el punto anterior será llamado *Modelo final* y a este modelo se le harán las respectivas validaciones de supuestos para un modelo de regresión lineal múltiple.

A. Supuesto de No correlación en los errores

En el modelo clásico de regresión lineal supone que no existe auto correlación en los errores ε_i , simbólicamente se tiene:

$$E(\varepsilon_i \varepsilon_j) = 0 \quad \forall i \neq j \Leftrightarrow Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$$

Para validar este supuesto se emplea la prueba de Breusch-Godfrey, que permite contrastar la hipótesis de auto correlación de un orden mayor a uno.

La hipótesis nula y alterna vienen dadas por:

- *Ho: Los residuales del modelo no siguen un proceso auto-regresivo de orden i*
- *Ha: Los residuales del modelo siguen un proceso auto-regresivo de orden i*

Se realizó el test hasta orden 3, obteniendo los siguientes resultados:

Order	df	LM-test	p-value
order up to 1	1	2,2741	0,1316
order up to 2	2	3,5958	0,1656
order up to 3	3	3,924	0,2698

Cada uno de los p-value es mayor que un nivel de significancia del 5%, por lo que existe evidencia estadística para no rechazar la hipótesis nula, luego los residuales NO siguen un proceso auto-regresivo orden *i*.

B. Supuesto de Homocedasticidad

La hipótesis nula y alterna vienen dadas por:

- H_0 : Los residuales presentan una varianza constante a través del tiempo
- H_a : Los residuales no presentan una varianza constante a través del tiempo

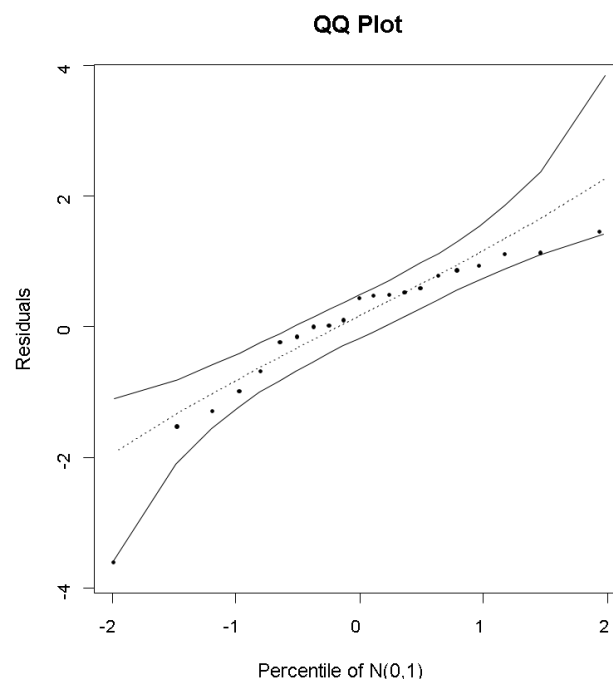
Según la literatura estadística, el supuesto de homogeneidad de varianza se puede validar por medio de la prueba de White, Breusch-Pagan-Godfrey, Harvey, Arch, entre otras. Para validar el supuesto se emplea la prueba de Breusch-Pagan-Godfrey, obteniendo los siguientes resultados:

Estadístico BP	Df	p-value
1,2694	2	0,5301

Se observa que hay suficiente evidencia estadística para no rechazar la hipótesis nula (valor $p = 0.5301 > 0.05$), en consecuencia, los residuales del modelo presentan una varianza constante, por lo que el supuesto se cumple estadísticamente.

C. Supuesto de Normalidad

Para la validación de este supuesto se utilizará un QQ-plot el cual comparará la distribución de los residuales del modelo con los de la distribución normal.



El gráfico anterior muestra sus respectivas bandas de confianza a un nivel del 95%, donde se evidencia que los puntos caen dentro de estas, razón por lo cual se puede afirmar que a un 5% de significancia los residuales presentan una distribución normal.

3.5. Final Model

El *modelo final* corresponde a un modelo estimado bajo una regresión lineal múltiple, el cual pretende estimar el recaudo de los impuestos prediales del municipio de Rionegro, para posteriormente poder realizar una estimación bajo el modelo.

El modelo es el siguiente:

$$\text{Recaudo} = 0,30228167 \cdot \text{IPVU_nom_medellin} - 5,68723669 \cdot \text{Tasa_ocupacion_7a} + 0,06976344 \cdot \text{Cambio_TRM}$$

Variable	Estimate	t -value	p-value
IPVU_nom_medellin	0,30228	17,277	0,000000
tasa_ocupacion_7a	-5,68724	-8,562	0,000000
cambio_trm	0,06976	3,983	0,000872
Residual standard error: 36.32 on 18 degrees of freedom Multiple R-squared: 0.9829, Adjusted R-squared: 0.9801 F-statistic: 345.2 on 3 and 18 DF, p-value: 4.341e-16			

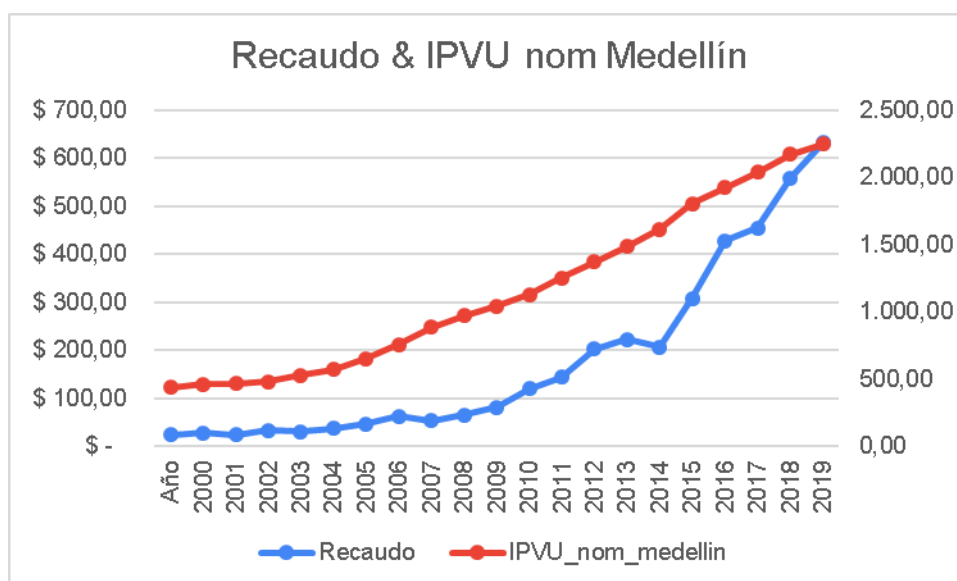
Donde,

- **Recaudo:** Recaudo anual de los ingresos por impuesto predial en el municipio de Rionegro dividido entre \$1.000.000.000

Las variables independientes del modelo son las siguientes:

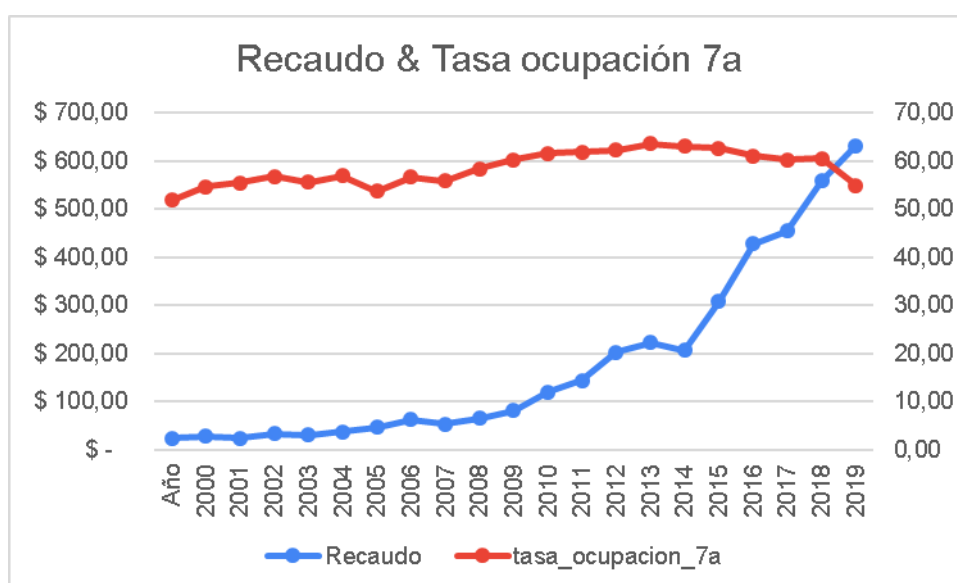
- **IPVU_nom_medellin:** Índice nominal de precios de vivienda usada para la ciudad de Medellín (Incluye Envigado, Bello e Itagüí).

Mide la evolución de los índices de precios de vivienda usada contemplando el efecto de la inflación, mediante su variación promedio para el periodo de análisis de forma anual. La metodología de cálculo de este indicador se encuentra publicada en *Borradores de Economía*, num. 368, de febrero de 2006 por el Banco de la república.



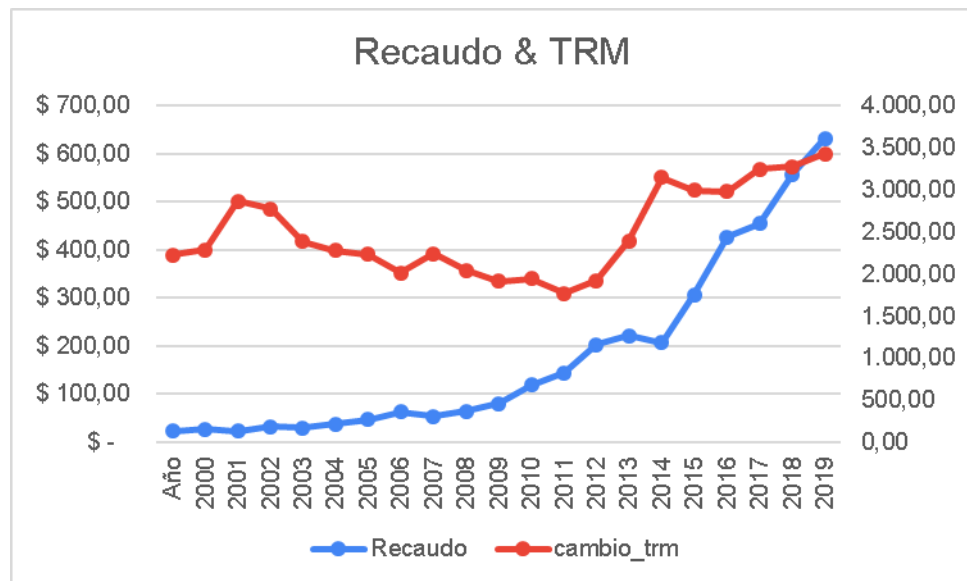
- **Tasa_ocupacion_7a:** Tasa de ocupación en las 7 áreas metropolitanas de Colombia a cierre de cada año

Relación porcentual entre la población ocupada y la población en edad de trabajar con base en la población de Bogotá, Barranquilla, Cali, Medellín, Bucaramanga, Manizales y Pasto. Este indicador es calculado por el Banco de la República basado en información de la GEIH (Gran Encuesta Integrada de Hogares) publicada por el DANE.



- **Cambio_trm:** Tasa de cambio representativa del mercado a fin de diciembre de cada año.

Cantidad de pesos colombianos por un dólar de Estados Unidos, la cual se calcula con base en las operaciones de compra y venta de divisas entre intermediarios financieros que transan en el mercado cambiario colombiano, con cumplimiento al mismo día cuando se realiza la negociación de las divisas. Este indicador es calculado por la Superintendencia Financiera de Colombia.

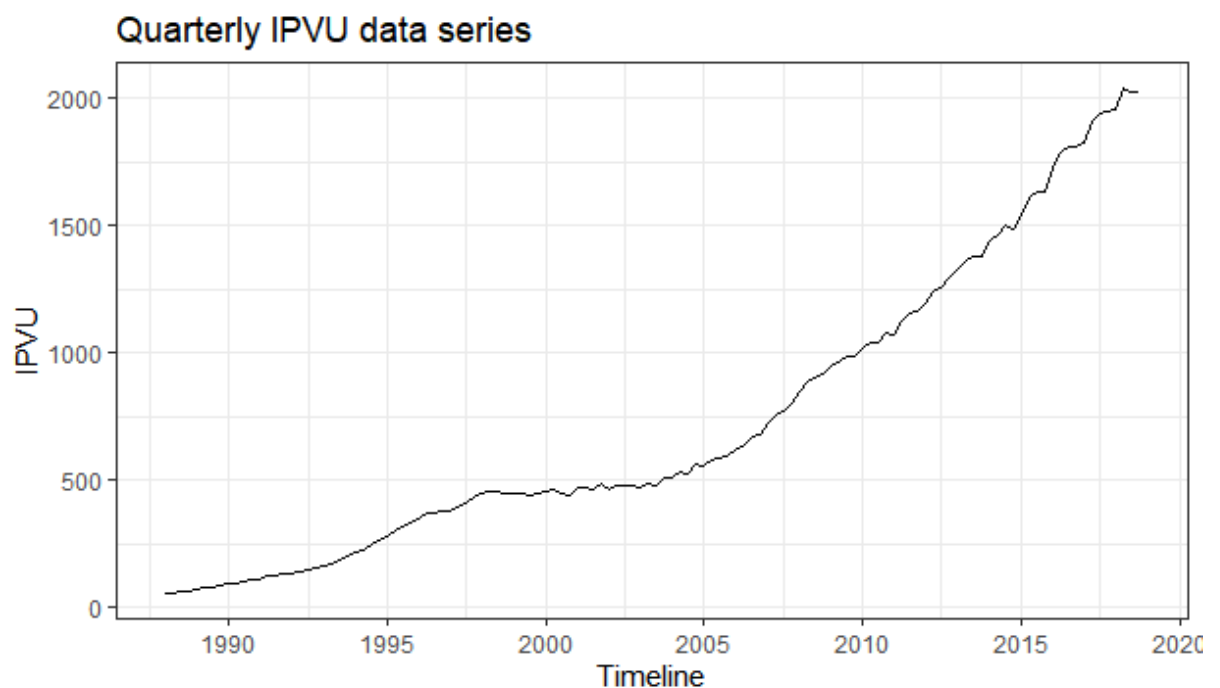


3.6. Forecast

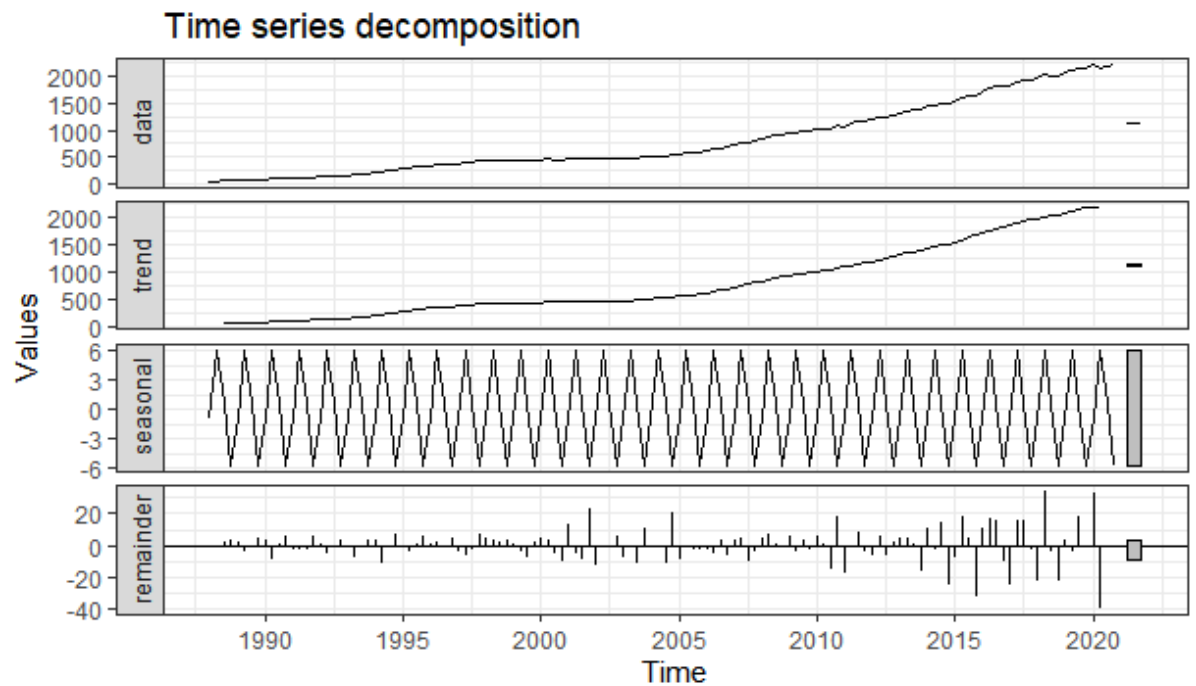
Para realizar el pronóstico de los siguientes dos años (2021 y 2022) se deben tener las predicciones de las variables dependientes del modelo final en estos dos periodos de tiempo. Se utilizaron diferentes métodos para obtener los pronósticos de las tres variables.

A. Forecast IPVU_nom_medellin

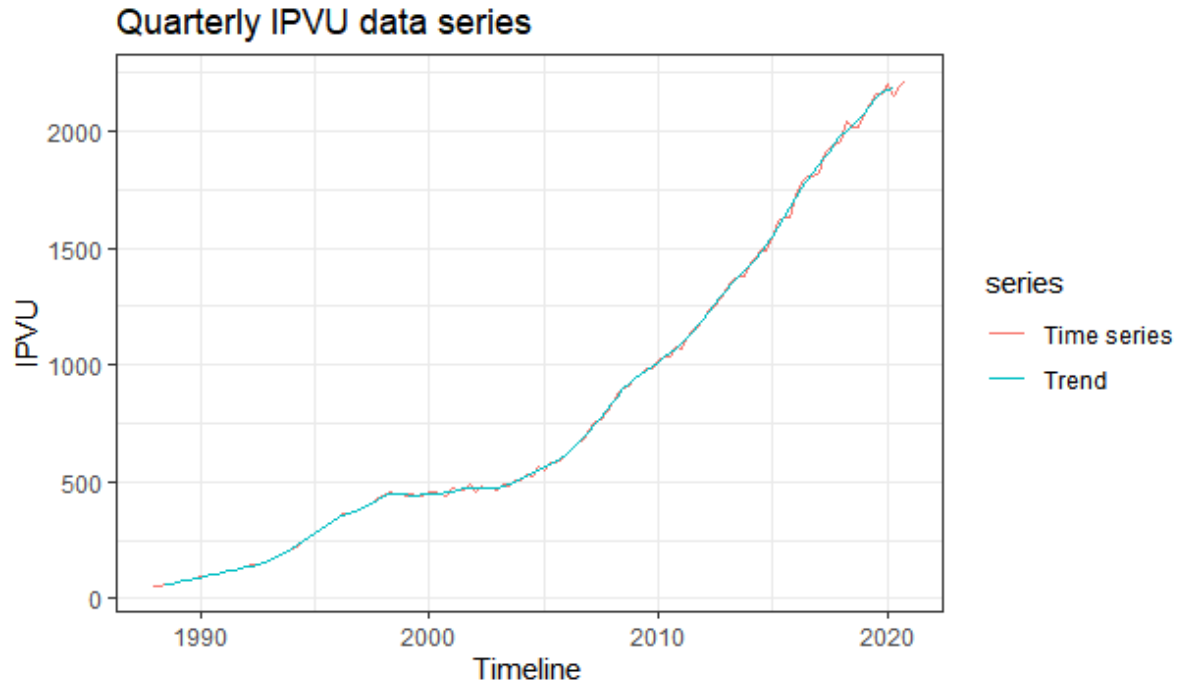
Se utilizaron los datos históricos del indicador trimestral suministrado por el Banco de la República, considerando como periodo inicial el primer trimestre del 1988 y como dato final el último trimestre del año 2020.

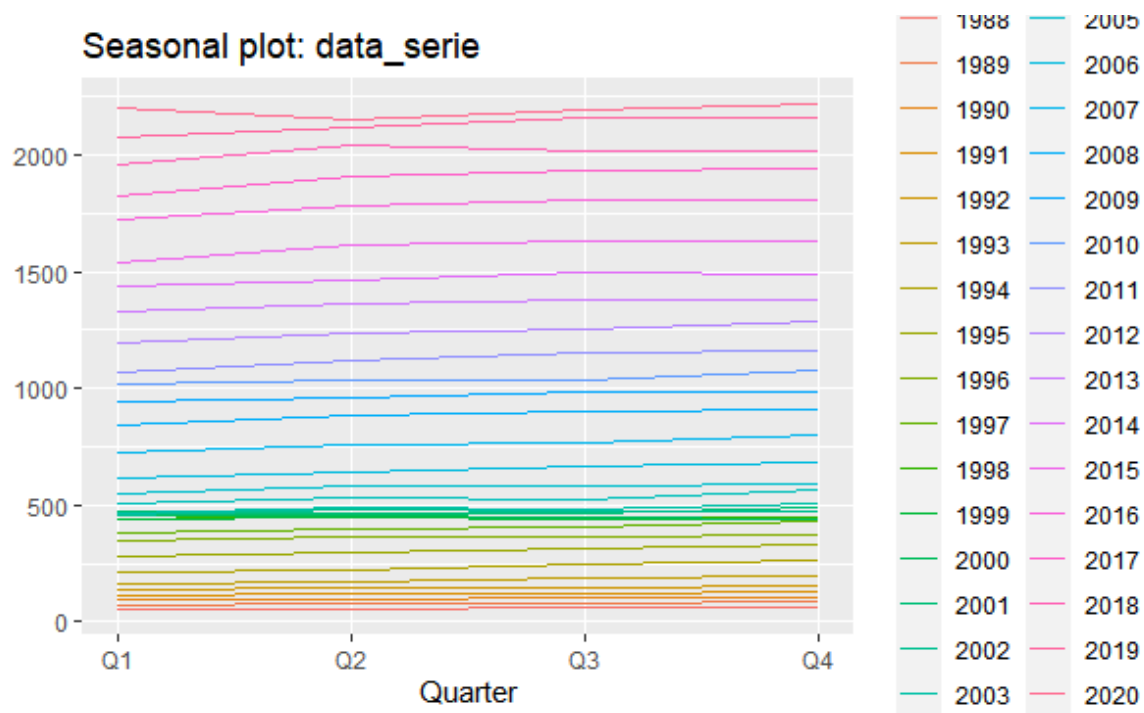


Se realizó la descomposición para analizar la tendencia y la estacionalidad de la serie:



En el gráfico anterior se muestra la evidencia de tendencia y estacionalidad de la serie, por lo cual se analizaron cada una de estas por aparte.

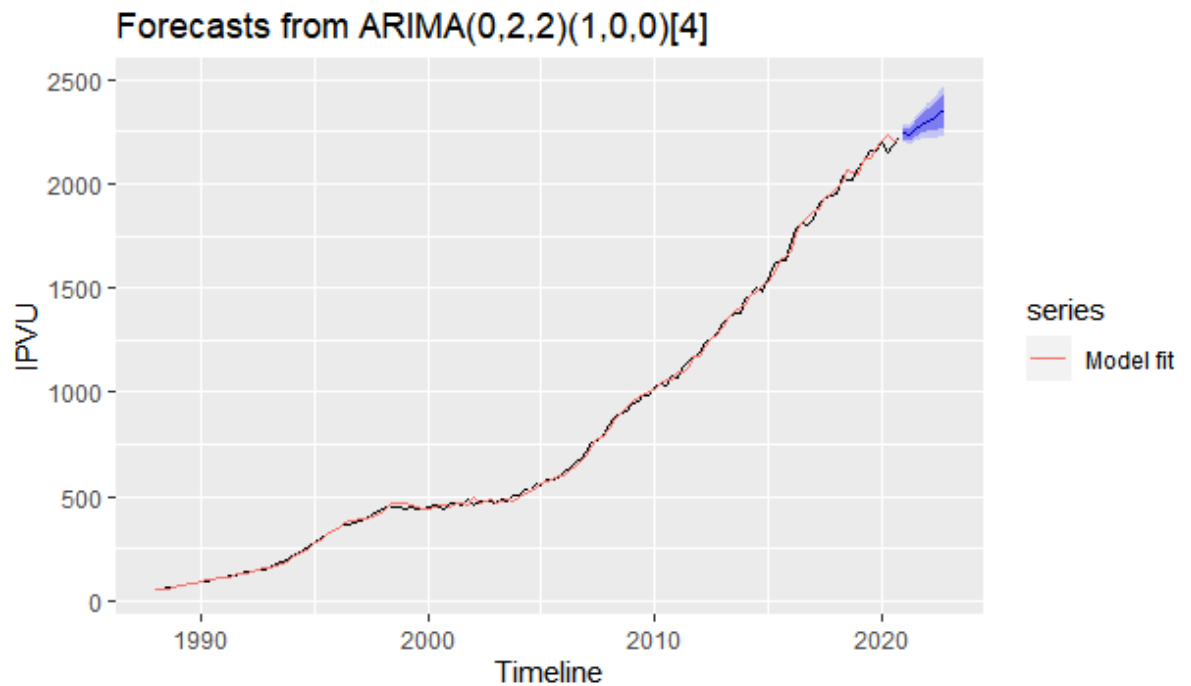




En las gráficas anteriores se muestra una clara tendencia de los datos a través del tiempo y una estacionalidad tomando una frecuencia trimestral.

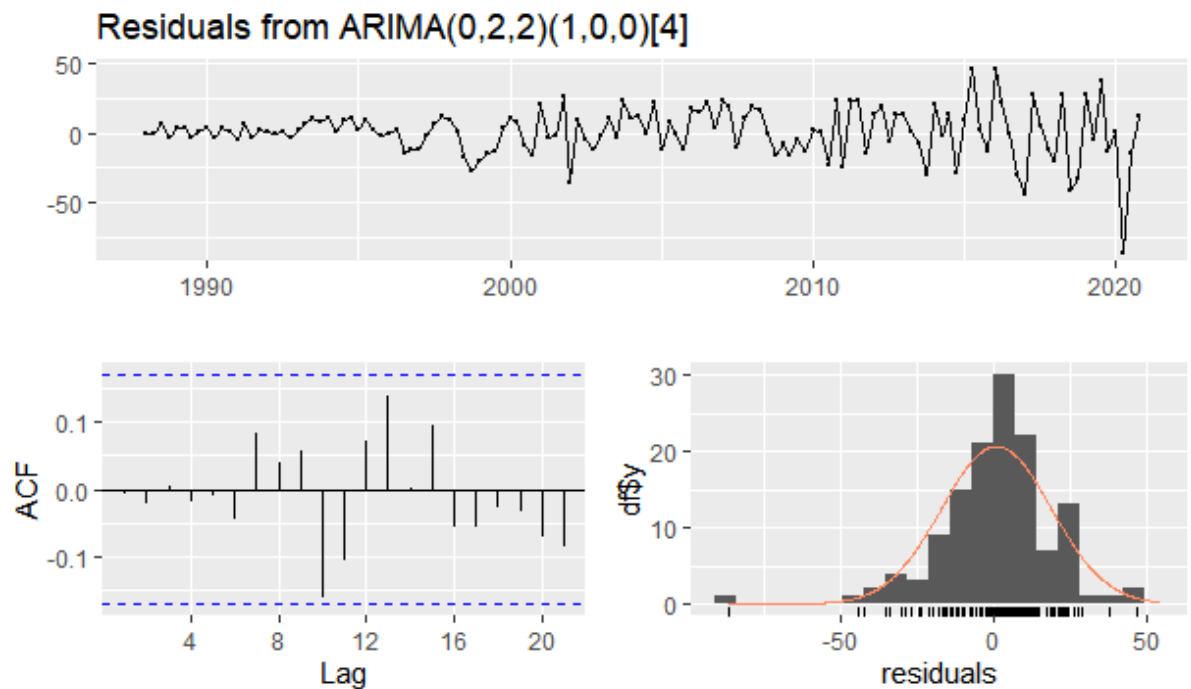
Teniendo en cuenta el análisis anterior se utilizó para el pronóstico un modelo $ARIMA(0,2,2)(1,0,0)$ con los siguientes parámetros:

Coefficients	ma1	ma2	sar1
	-1,2191	0,3041	0,3724
estimation series	0,0856	0,0933	0,0993



En la gráfica anterior se muestra la estimación del modelo (línea roja), los datos reales (línea negra) y el pronóstico de los siguientes 2 años (8 puntos trimestrales).

Por último se realizó un análisis sobre los residuales donde se validó que su distribución se asemejara a una distribución normal.



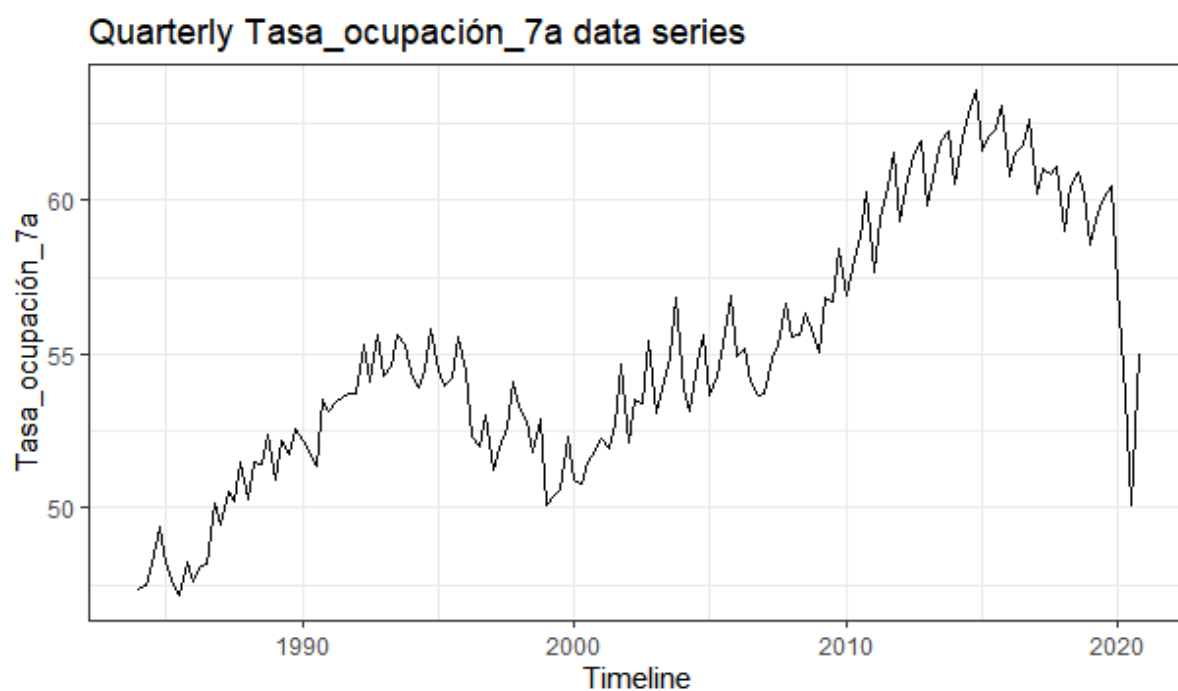
Con el modelo anterior se realizaron pronósticos trimestrales para un año con un intervalo de confianza del 95%.

	2021Q1	2021Q2	2021Q3	2021Q4	Average
--	--------	--------	--------	--------	---------

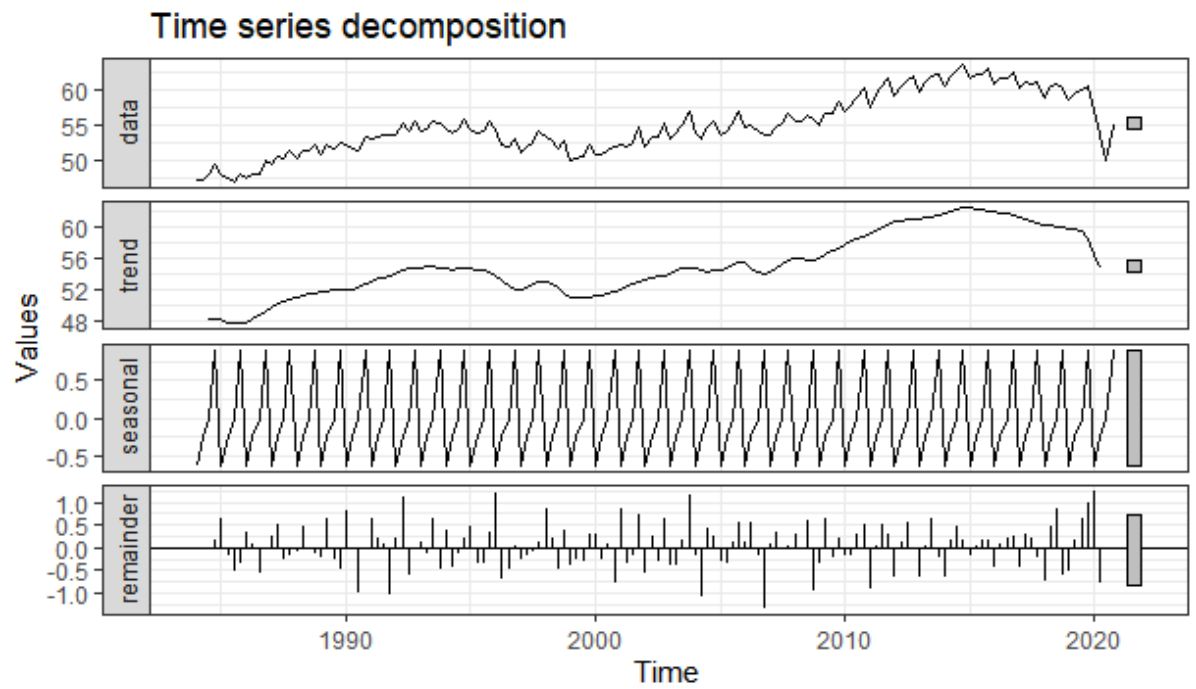
CI lower	2207,889	2189,782	2207,425	2219,236	2206,083
CI mean	2243,436	2234,883	2262,028	2283,455	2255,9505
CI upper	2278,983	2279,985	2316,632	2347,674	2305,8185

B. Forecast Tasa_ocupacion_7a

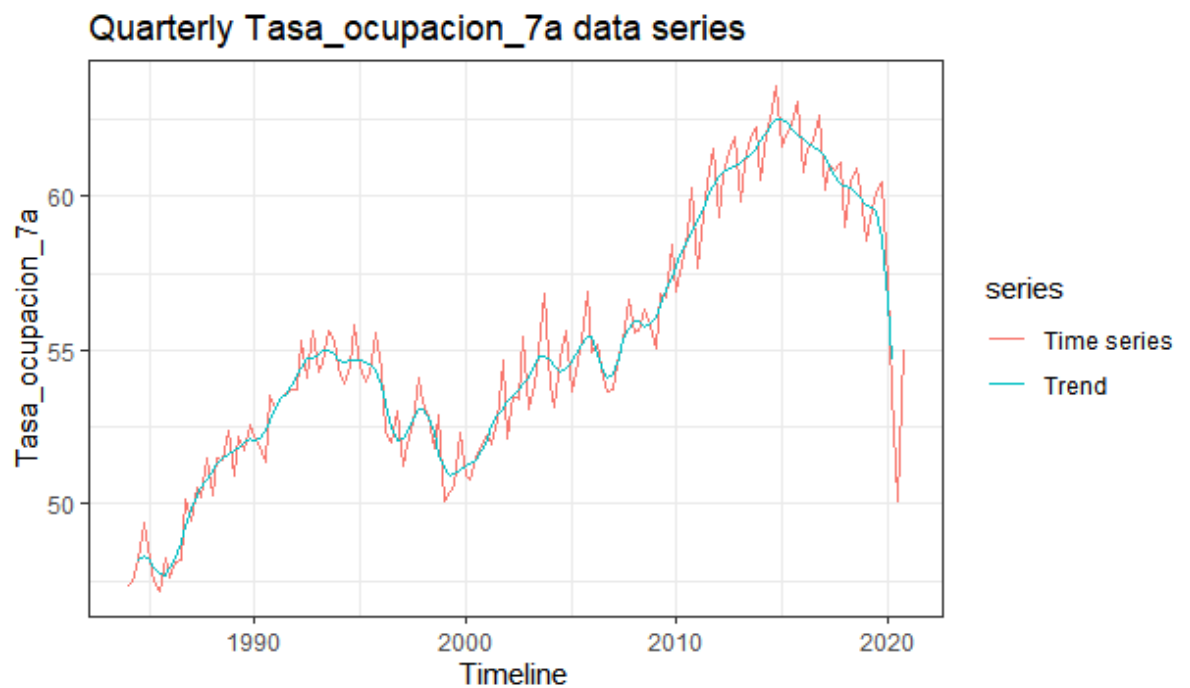
Se utilizaron los datos históricos del indicador trimestral suministrado por el Banco de la República, considerando como periodo inicial el primer trimestre del 1984 y como dato final el último trimestre del año 2020.

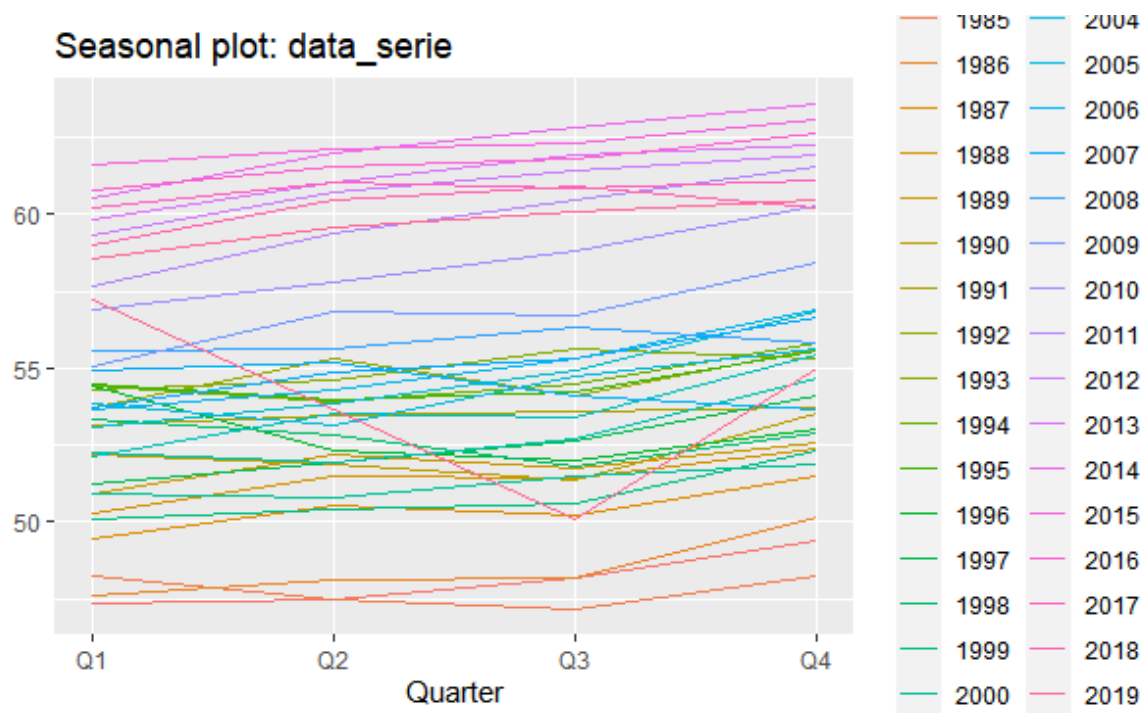


Se realizó la descomposición para analizar la tendencia y la estacionalidad de la serie:



En el gráfico anterior se muestra la evidencia de tendencia y estacionalidad de la serie, por lo cual se analizaron cada una de estas por aparte.

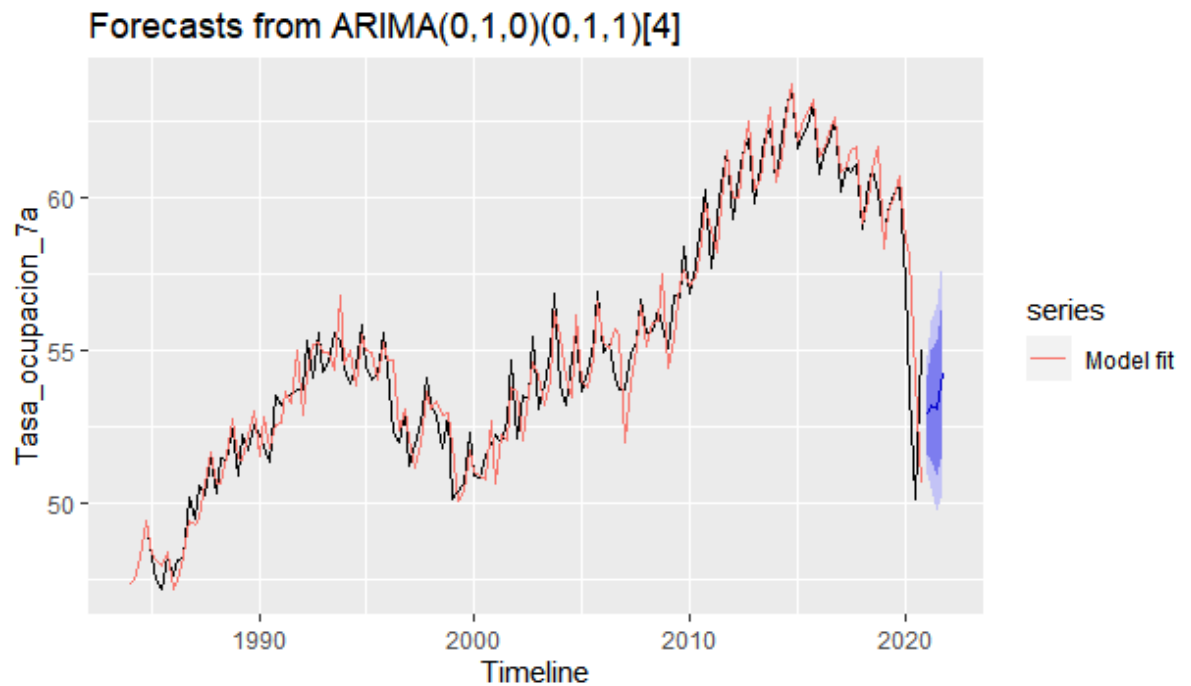




En las gráficas anteriores se muestra una clara tendencia de los datos a través del tiempo pero se observa que no hay estacionalidad tomando una frecuencia trimestral.

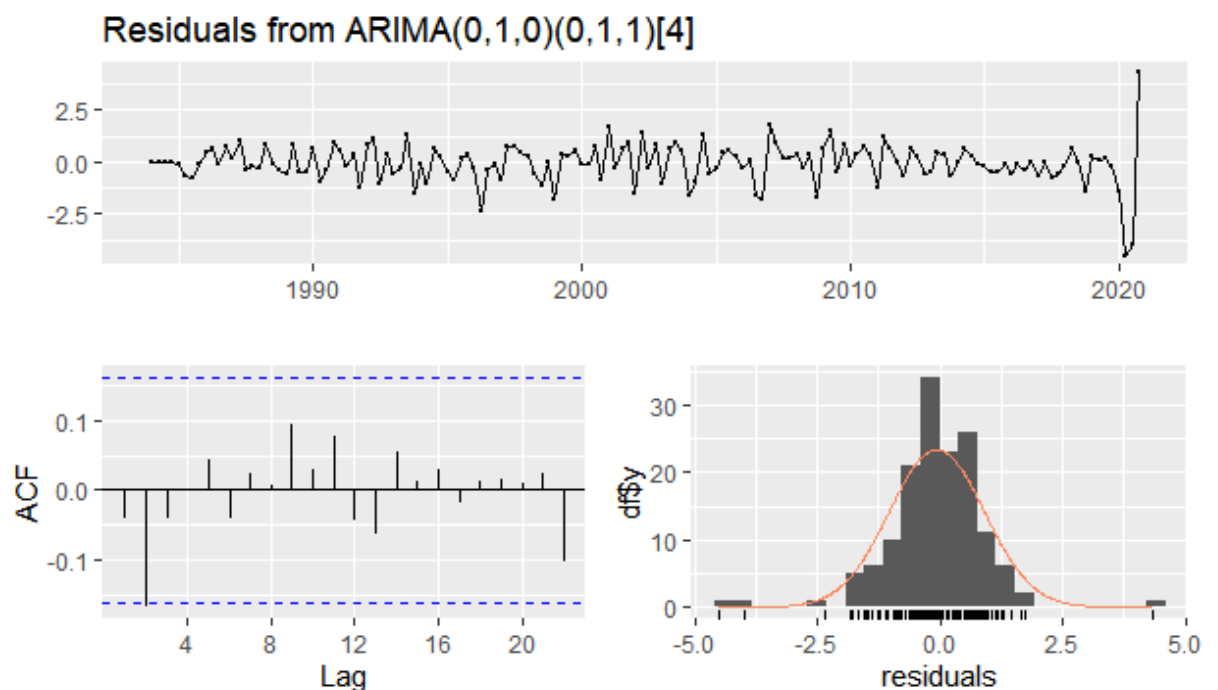
Teniendo en cuenta el análisis anterior se utilizó para el pronóstico un modelo $ARIMA(0,1,0)(0,1,1)$ con los siguientes parámetros:

Coefficients	sma1
	-0,8606
estimation series	0,0728



En la gráfica anterior se muestra la estimación del modelo (línea roja), los datos reales (línea negra) y el pronóstico de los siguientes 2 años (8 puntos trimestrales).

Por último se realizó un análisis sobre los residuales donde se validó que su distribución se asemejara a una distribución normal.



Con el modelo anterior se realizaron pronósticos trimestrales para un año con un intervalo de confianza del 95%.

	2021Q1	2021Q2	2021Q3	2021Q4	Average
CI lower	50,9457	50,46851	49,74392	50,37086	50,3822475
CI mean	52,88189	53,20669	53,09749	54,24323	53,357325
CI upper	54,81808	55,94488	56,45106	56,77524	55,997315

C. Forecast Cambio_TRM

Para el pronóstico de esta variable se recomienda enfocarse en los pronósticos mostrados en la siguiente tabla los cuales fueron tomados de la encuesta electrónica del Banco de la república:

Tasa de Cambio Nominal		
Analistas Locales		
	Alianza Valores	3.750
	ANIF	n.d.
	Banco de Bogotá	3.950
	Bancolombia	n.d.
	BBVA Colombia	3.430
	BTG Pactual	n.d.
	Corficolombiana	3.557
	Corredores Davivienda /2	n.d.
	Credicorp Capital /3	3.300
	Davivienda	n.d.
	Fedesarrollo	n.d.
	Itaú /1	3.950
	Ultraserfinco /4	n.d.
	Promedio	3.656
Analistas Externos		
	Citigroup	n.d.
	Deutsche Bank	n.d.
	Goldman Sachs	3.200
	JP Morgan	n.d.
	Promedio	3.200
/1. Antiguo Corpbanca, hasta junio de 2017.		
/2. Antiguo Corredores Asociados		
/3. Antiguo Correval		

/4. Antiguo Ultrabursátiles
n.d.: no disponible
Fuente: Banco de la República (encuesta electrónica)


4. Web application

Front End Mockups

We present two pages that pretend to have a friendly interaction with the user. The web application was thought as a system, in such a manner that a set of inputs is expected and a couple of outputs will be returned such as a time line plot and card showing a number.

Even if we suggest the value of inputs the user will be able to modify the values and see how the forecast performs, we also aim to enable a space for uploading files in order to eventually re-train the models and take more information in consideration.


In conclusion, the web application will work as a tool where Rionegro's town hall will be able to make a forecast of the billing of predial taxes.



Rio Analytics


Please enter the following data

PREDICT

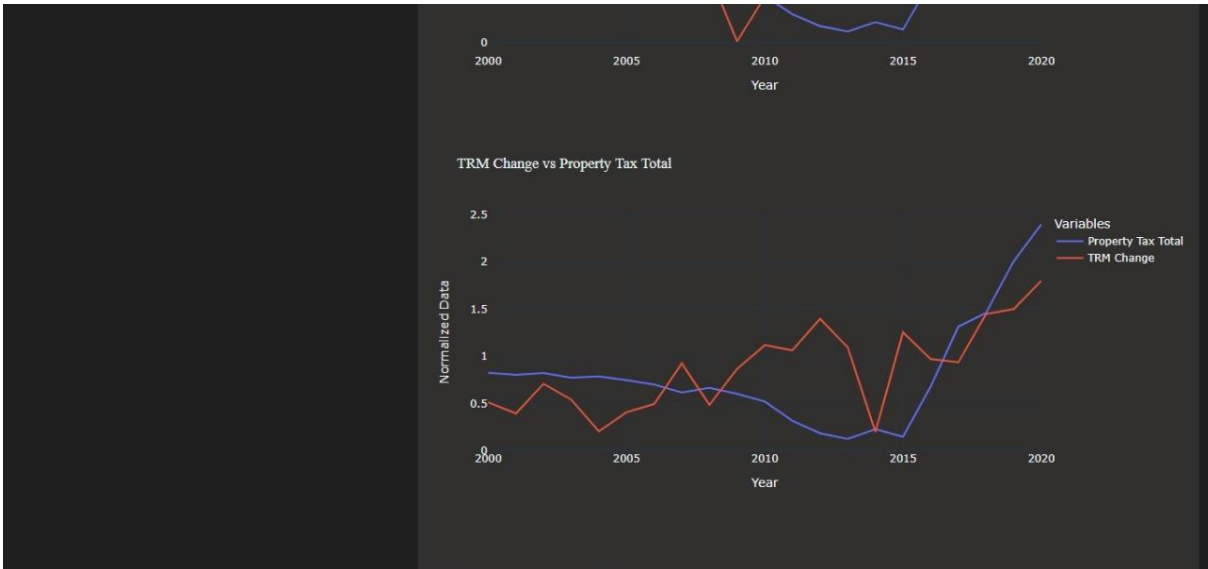
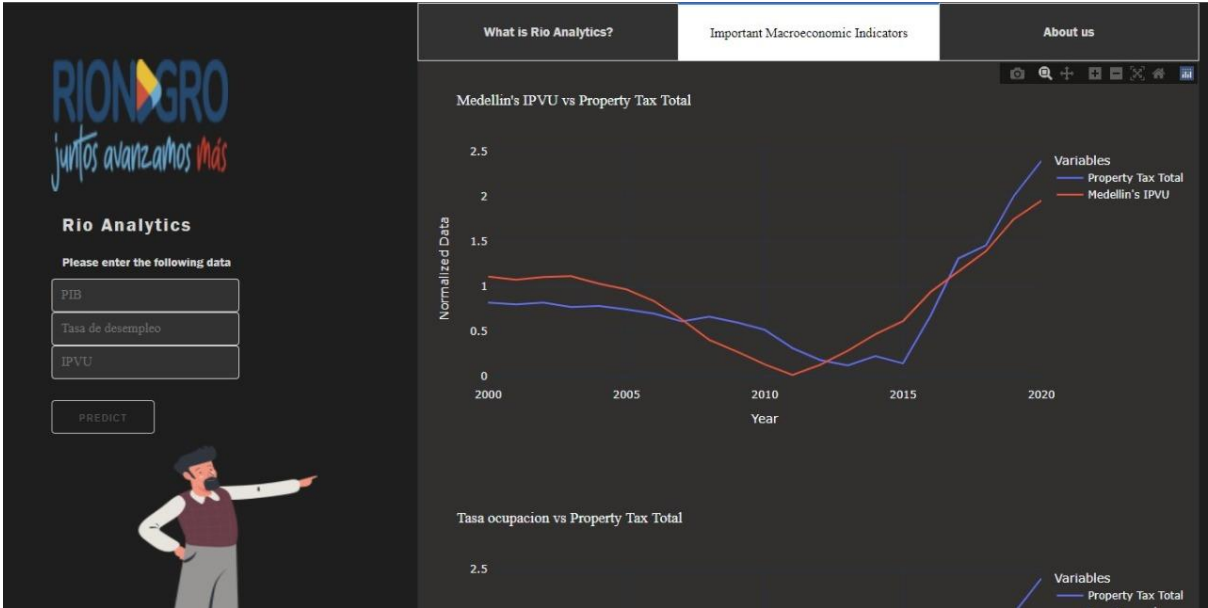


What is Rio Analytics?
Important Macroeconomic Indicators
About us

Rio Analytics



Rionegro's Town (one of Antioquia's department municipalities) recognizes and suffers from the fact that sources of public resources are limited and the social needs tend to increase over time. Therefore, one of their aims is to predict incomes and then use this information to optimize the short and long term expenditure & investment planning and execution processes and thus solve the problem of planning and making investments without knowing what the actual income will be. An important source of this income is property tax. Project scope is limited to predict this specific income With optimal prediction of income from property taxes, among others, Rionegro's town will be able to plan and manage in advance their financial resources and direct and lead them in an efficient way to address their social investment and expenditure needs. The correct planning and management of the resources, investments and expenditure helps to meet and address population and community needs.



5. Conclusions and future work

Rionegro as all municipalities around the country recognize and suffer from the fact that sources of public resources are limited and the social needs tend to increase over time. Through the paper we evidenced clearly the importance of making accurate forecast of the income and we

Number of properties, total constructed areas, and total amount to be paid have been increasing over time, so we found an upward trend that a priori gave us a perspective of where income will be. We also found that the largest properties are located in high strata and also were identified destinations where the average invoice was higher than the others. Important and expected correlations among variables were confirmed like total constructed area and total value of the properties and the tax to be paid as expected is highly correlated with rate and the property appraisal.

include macroeconomics indexes was a success because it helped to understand trends and finally IPVU....

For future releases we plan to make a segmentation model and apply custom models to each group in order to increase accuracy in the final result. Group properties would help to identify different trends and predict how sectors will evolve over time, keeping this in mind we could implement specific strategies in each different group.