Presentación del curso

Roberto Mendoza Matos

March 22, 2023

Primera parte

- Objetos en R y Python (listas, vectores, diccionarios, tuplas)
- Funciones, loops. Funciones del tipo apply.
- Estructuras de Clase en Python
- Limpieza de base de datos (merge, append, pivot, reshape)
- Ejemplos en ENAHO, ENDES y CE
- Extraer información de PDF

Expresiones regulares

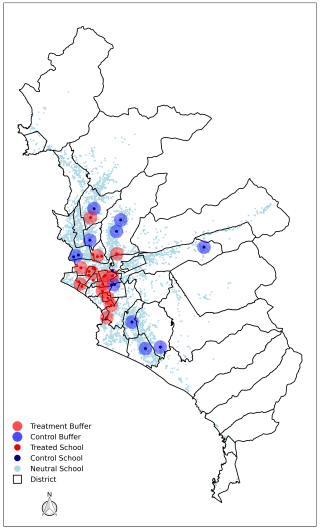
- Permite detectar patrones en texto
- 50% de Text mining se resuelve con REGEX

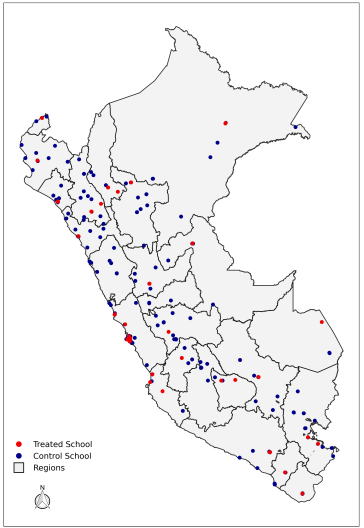
```
index_columns_traffic = np.where( data.columns.str.contains('traffic$', regex=True))[0]
index_columns_metro = np.where( data.columns.str.contains('metro$', regex=True))[0]
index_columns_traffic_pr = np.where( data.columns.str.contains('^priv.*traffic$', regex=True))[0]
index_columns_metro_pr = np.where( data.columns.str.contains('^priv.*metro$', regex=True))[0]
index_columns_traffic_pub = np.where( data.columns.str.contains('^pub.*traffic$', regex=True))[0]
index_columns_metro_pub = np.where( data.columns.str.contains('^pub.*metro$', regex=True))[0]
index_columns_traffic_elite = np.where( data.columns.str.contains('^elite.*traffic$', regex=True))[0]
index_columns_metro_elite = np.where( data.columns.str.contains('^elite.*metro$', regex=True))[0]
```

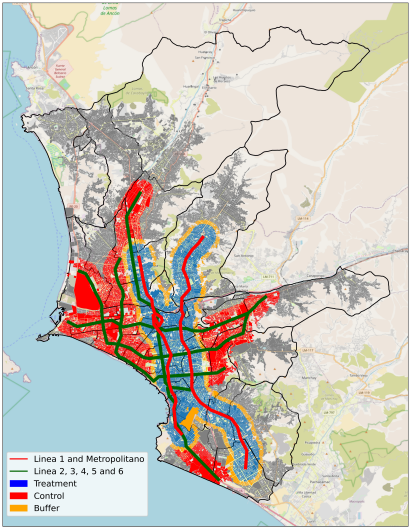
- (REGEX link)

Geopandas

- Mapas, buffers, Points

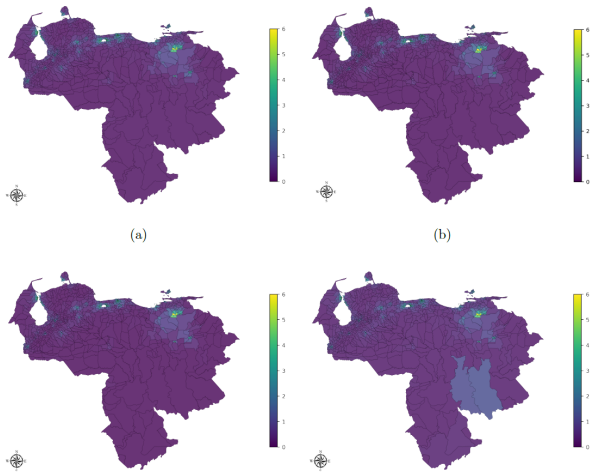






Geopandas

- Mapas de Calor



(a) (b)

Tables en Latex

- Estadísticas descriptivas

Table 1: Descriptive Statistics (Treatment group)

Statistic	N	Mean	Median	St. Dev.	Min	Max
Average score	8,991	13.74	13.55	1.37	9.27	19.64
Math score	8,991	14.04	14	1.74	4	20
Lenguage score	8,991	13.25	13	1.87	7	20
IFH-score	8,863	0.28	−0.11	0.99	−2.33	4.24
Years of education	9,114	7.55	8	1.45	1	13
Head household's education in year	7,648	5.84	6	3.02	1	17
No extreme poverty	8,832	0.47	0	0.50	0	1
Extreme poverty	8,832	0.37	0	0.48	0	1
Age	9,120	15.23	15	1.20	14	19
Gender	9,120	0.50	0	0.50	0	1
Goods score	6,497	−0.09	−0.11	0.12	−0.11	1.09
Electrivity access score	6,497	0.09	0.22	0.22	−0.29	0.22
Floor score	6,488	−0.14	−0.17	0.12	−0.17	0.47
Roof score	6,497	0.00	0	0.00	0	0
Maximun Educational attaiment score	6,497	0.39	0.41	0.10	−0.35	0.83

Source: SISFOH (2013) and SIAGIE 2014

Tables en Latex

- Estadísticas descriptivas

Table 1: Nighlight Venezuela

	N	Mean	Median	SD	Min	Max
2013	1134	9.04	0.72	53.06	0.03	1444.65
2014	1134	9.68	0.67	60.40	-0.02	1407.17
2015	1134	8.59	0.57	58.98	0.02	1517.33
2016	1134	7.55	0.50	55.47	-0.03	1564.49
2017	1134	6.64	0.64	38.31	0.15	1102.76
2018	1134	6.50	0.60	41.38	0.16	1119.54
2019	1134	7.00	0.53	56.97	0.11	1671.13
2020	1134	5.91	0.59	34.68	0.18	1010.58

Table 2: Nighlight Venezuela (inverse hyperbolic transformation)

	N	Mean	Median	SD	Min	Max
2013	1134	1.00	0.47	1.20	0.03	5.41
2014	1134	0.98	0.43	1.23	-0.01	5.94
2015	1134	0.92	0.40	1.17	0.02	5.91
2016	1134	0.85	0.34	1.15	-0.03	5.87
2017	1134	0.97	0.49	1.08	0.15	5.47
2018	1134	0.94	0.47	1.04	0.16	5.58
2019	1134	0.89	0.43	1.04	0.11	5.73
2020	1134	0.94	0.49	1.02	0.18	5.24

Tables en Latex

-Cuadros de regresiones

Table 1: Potato Market modelling (Linear regression)

Dependent Variable: Potato price (log)				
	Model 1	Model 2	Model 3	Model 4
Potato production (log)	-0.002	0.003***	-0.001	-0.003***
	(0.000)	(0.001)	(0.001)	(0.001)
Dummy April		-0.057***	-0.022***	-0.024***
		(0.009)	(0.002)	(0.004)
Dummy May		-0.070***	-0.029***	-0.029***
		(0.014)	(0.003)	(0.003)
Dummy June		-0.042***	-0.002**	-0.002
		(0.015)	(0.001)	(0.004)
Month fixed effects	no	yes	yes	yes
Year fixed effects	no	no	yes	yes
Province fixed effects	no	no	no	yes
Observarions	2911	2911	2911	2911
R^2	0.0	0.027	0.647	0.650

Note.Huber robust standard errors are in parentheses.
Regression disturbance terms are clustered at the District level.
Rest of dummies by month, year and province fixed effects not reported
* Significantly different from zero at 90 percent confidence.
** Significantly different from zero at 95 percent confidence.
*** Significantly different from zero at 99 percent confidence.

	Math score		
	(1)	(2)	(3)
Variables			
Grade score	-0.0317	-0.0422**	-0.0479**
	(0.0195)	(0.0170)	(0.0169)
Observations	6,228	5,199	5,199
Control variables	No	Yes	Yes
Fixed effects	No	No	Yes
Dependent mean	2.755	2.795	2.795

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Fixed effects by district and control variables not reported

	Language score		
	(1)	(2)	(3)
Variables			
Grade score	0.0877***	0.0950***	0.0886***
	(0.0249)	(0.0279)	(0.0274)
Observations	6,228	5,199	5,199
Control variables	No	Yes	Yes
Fixed effects	No	No	Yes
Dependent mean	2.755	2.795	2.795

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Fixed effects by district and control variables not reported

Web scrapping

- MEF inversiones (MEF link)
- Consula Amigable (Consulta Amigable link)
- Mara de I.E del Censo Escolar (MINEDU) (MINEDU link)
- Contratación administrativo de servicios (CAS) (CAS link)

- Uso de la libreria Selenium (Disponible en R y Python)

Google Cloud

- Geocoding
- Travel time
- Vision API

Geocoding

Entregamos direcciones y obtenemos las coordenadas

- Censo universitario (college student address)
- Registro de hechos delictivos (INEI)

```
df.head()
```

	Address	geocoded
0	Third Floor, Century City Mall, Kalayaan Avenu...	(14.565466, 121.0276651)
1	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	(14.5535301, 121.0146986)
2	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	(14.5813906, 121.0570799)
3	Third Floor, Mega Fashion Hall, SM Megamall, O...	(14.5844135, 121.0566779)
4	Third Floor, Mega Atrium, SM Megamall, Ortigas...	(14.5846851, 121.0573023)

Travel Time

```
1 data_avenues
```

street	origin	destination	speed_pess	obs_pess	speed_best	obs_best	speed_opt	obs_opt
Avenida Paseo de la República	-12.132344368406516,-77.02193718479606	-12.091300521004715,-77.02274800315827	37.929760	Traffic Speed	52.480818	Traffic Speed	60.530973	Traffic Speed
Vía Expresa Luis Fernán Bedoya Reyes	-12.08959190612168,-77.02315481504564	-12.063076324350194,-77.0341029591271	60.314136	Traffic Speed	68.165680	Traffic Speed	83.478261	Traffic Speed
Vía Expresa Grau	-12.057049394213886,-77.01540445511357	-12.059557340809123,-77.03498815015124	17.428139	Traffic Speed	28.444444	Traffic Speed	28.944724	Traffic Speed
Plaza Francisco Bolognesi	-12.05981066744398,-77.03516948066084	-12.059911998030914,-77.03791534266348	13.104925	Traffic Speed	22.014388	Traffic Speed	20.675676	Traffic Speed
Avenida 9 de Diciembre	-12.059861332742225,-77.03799305573902	-12.060190656947565,-77.0409720569683	17.142857	Traffic Speed	25.714286	Traffic Speed	33.600000	Traffic Speed
Plaza Francisco Bolognesi	-12.060139991711532,-77.04094615260979	-12.058138707231999,-77.04169737927761	15.000000	Traffic Speed	22.222222	Traffic Speed	23.529412	Traffic Speed
Avenida Alfonso Ugarte	-12.057961377507253,-77.04177509235315	-12.04068850403358,-77.0433785283332	23.259912	Traffic Speed	37.894737	Traffic Speed	41.357702	Traffic Speed
Avenida Caquetá	-12.035854591489796,-77.04359342165324	-12.028645453515216,-77.04451480628498	20.065574	Traffic Speed	25.606695	Traffic Speed	24.189723	Traffic Speed
Avenida Túpac Amaru	-12.028244988630773,-77.04456184574795	-11.984873364349381,-77.05781153539735	25.116279	Traffic Speed	41.707317	Traffic Speed	47.610209	Traffic Speed
Avenida Alfredo	-12.128778766939579,-77.00180372445496	-12.12542634930522,-77.02308072260814	19.211137	Traffic	31.011236	Traffic	39.428571	Traffic

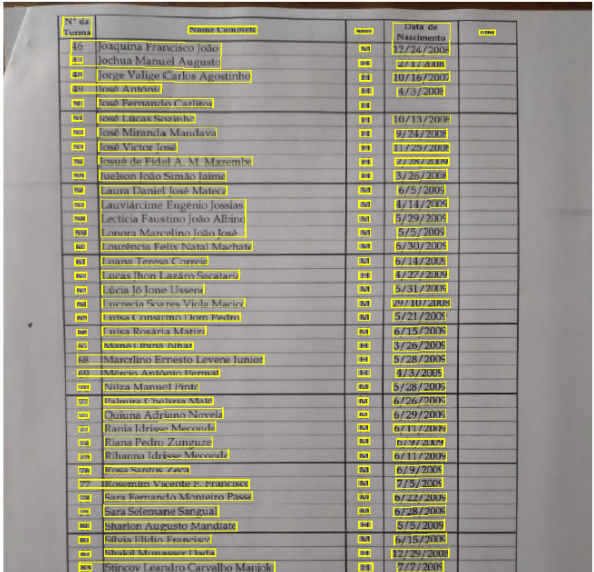
Travel Time

A partir de direcciones o coordenadas de origen y destino se puede obtener:

- Tiempo de tranporte (private driving, walking and public transport)
- Diferentes escenarios (best, pessimistic and optimistic)
- Se puede hallar el travel time de años anteriores

Google Vision API

Reconocimiento de texto en handwriting o pdf scaneados:



Nº da Turma	Nome Completo		Data de Nascimento	
46	Joaquina Francisco João		12/24/2008	
47	Joshua Manuel Augusto		[illegible]	
48	Jorge Valige Carlos Agostinho		10/16/2007	
49	José Antonio		4/3/2008	
50	José Fernando Carlitos			
51	José Lucas Sesinho		10/15/2008	
52	José Miranda Mandava		9/24/2008	
53	José Victor José		11/25/2008	
54	Josué de Fidel A. M. Mazembe		[illegible]	
55	Joelson João Simão Jaime		3/26/2008	
56	Laura Daniel José Mateo		6/5/2005	
57	Lauviárcine Eugénio Jossias		4/14/2009	
58	Lecticia Faustino João Albino		5/29/2006	
59	Leonora Marcelino João José		5/5/2006	
60	Laurência Felix Natal Machate		6/30/2006	
61	Luana Teresa Correia		6/14/2006	
62	Lucas Ihon Lazáro Secatari		4/27/2009	
63	Lúcia Jó Jone Ussene		5/31/2006	
64	Lucrecia Soares Viola Maciel		[illegible]	
65	Luisa Consumo Dom Pedro		5/21/2005	
66	Luisa Rosária Mario		6/15/2006	
67	Manoel Dutra Silva		3/26/2006	
68	Marcelino Ernesto Levene Junior		5/28/2006	
69	Marcio Antonio Bernal		4/3/2005	
70	Nilza Manuel Pinto		5/28/2006	
71	Paloma Chaloma Mate		6/26/2006	
72	Quiuna Adriano Novela		6/29/2006	
73	Rania Idrisse Mecondi		6/11/2009	
74	Riana Pedro Zunguze		[illegible]	
75	Rihanna Idrisse Mecondi		6/11/2009	
76	Rosa Santos Zena		6/9/2004	
77	Rosemim Vicente F. Francisco		7/5/2006	
78	Sara Fernando Monteiro Passe		6/22/2006	
79	Sara Selemane Sargual		6/28/2006	
80	Sharlon Augusto Mandlate		5/5/2009	
81	Silvia Flávio Francisco		6/15/2006	
82	Shakil Mamassse Charfa		12/29/2008	
83	Stincov Leandro Carvalho Manjolo		7/7/2005	

Google Vision API

Se obtiene cada palabra en un rectangulo de coordenadas

```
In [131]: 1 Outcome.text_detec()
text_annotations {
  description: "Rosaria"
  bounding_poly {
    vertices {
      x: 313
      y: 1008
    }
    vertices {
      x: 352
      y: 1008
    }
    vertices {
      x: 352
      y: 1018
    }
    vertices {
      x: 313
      y: 1018
    }
  }
In [132]: 1 Outcome.text_detec()
    }
  }
  text_annotations {
    description: "6/9/2009"
    bounding_poly {
      vertices {
        x: 522
        y: 577
      }
      vertices {
        x: 584
        y: 577
      }
      vertices {
        x: 584
        y: 588
      }
      vertices {
        x: 522
```

Sacarse un 10

Calificacion

NO HAY EXAMEN PARCIAL

VII. EVALUACIÓN

TIPO DE EVALUACIÓN	CANTIDAD	FECHA(S)	PESO TOTAL (EN %)
Trabajos grupales semanales (TG)	7*	8 de abril 15 de abril 22 de abril 29 de abril 6 de mayo 3 de junio 10 de junio	40%
Participación (PR)	1		20%
Trabajo Final (TF)	1	16 de Julio	40%

** Se eliminará la nota más baja.*

Trabajo Grupal

- Grupos libres de 3 integrantes **(Link de grupos en PAIDEIA)**
- Envió por correo a los integrantes del grupo de 4 personas
- Tareas semanales (R y Python)
- Entrega: sábados 11:59 pm via Git Hub
- El orden, comentar sus códigos y que no haya falla al correr. Correr todo para revisar que no haya errores.

Trabajo Final

- Temas posibles (Regresiones, Web scrapping, Geocoding, Travel time)
- Asignación del trabajo final (12 de junio)
- Entrega 16 de julio (2° semana de exámenes finales)
- Adicional a los códigos se debe entrega del reporte mediante Overleaf.

Instalación

- Instalación de R y Python
- Tómese su tiempo para instalar los programas con paciencia
- El tutorial de instalación se enviará por correo

Entrega de trabajos y comunicación

- Instalar Slack (Chat de trabajo)
- Envío de capturas de tareas cortas en clase **(Nota de participación)**
- Crear cuenta en Git Hub. Indicar cuenta de GitHub y correo **(Link Paideia)**

Checklist

- Instalar R y Python
- Cuenta Git Hub (usar correo PUCP)
- Instalar Slack (usar correo PUCP)
- Crear cuenta overleaf (usar correo PUCP)
- Selección de la delegada o delegado