PARTE I. OUTLIERS Y MISSINGS

1. CONJUNTO DE DATOS

Tabla que posee información de las actividades de distintos números de teléfono, en este caso la variables objetivo indica True si el cliente con ese número telefónico cancelar su servicio y False si aún sigue en la compañia de telefonos.

Tabla "Títulos de Churn Telecom"

state	account length	area code	phone number	internati onal plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls	churn
KS	128	415	382-4657	no	yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	3	2.7	1	FALSE
OH	107	415	371-7191	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7	1	FALSE
NJ	137	415	358-1921	no	no	0	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29	0	FALSE
OH	84	408	375-9999	yes	no	0	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	FALSE
OK	75	415	330-6626	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	FALSE
AL	118	510	391-8027	yes	no	0	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.7	0	FALSE
MA	121	510	355-9993	no	yes	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	02.03	3	FALSE
MO	147	415	329-9001	yes	no	0	157	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92	0	FALSE
LA	117	408	335-4719	no	no	0	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35	1	FALSE
WV	141	415	330-8173	yes	yes	37	258.6	84	43.96	222	111	18.87	326.4	97	14.69	11.2	5	03.02	0	FALSE
IN	65	415	329-6603	no	no	0	129.1	137	21.95	228.5	83	19.42	208.8	111	9.4	12.7	6	3.43	4	TRUE
RI	74	415	344-9403	no	no	0	187.7	127	31.91	163.4	148	13.89	196	94	8.82	9.1	5	2.46	0	FALSE

State :	categorical, for the 50 states and the District of Columbia
Account Length :	integer-valued, how long account has been active
Area Code :	categorical
Phone :	Phone number of customer
Int'l Plan :	International plan activated (yes , no)
VMail Plan :	Voice Mail plan activated (yes , no)
VMail Message :	No. of voice mail messages
Day Mins :	Total day minutes used
Day Calls :	Total day calls made
Day Charge :	Total day charge
Eve Mins :	Total evening minutes
Eve Calls :	Total evening calls
Eve Charge :	Total evening charge
Night Mins :	Total night minutes
Night Calls :	Total night calls
Night Charge :	Total night charge
Intl Mins :	Total International minutes used
Intl Calls :	Total International calls made
Intl Charge :	Total International charge
CustServ Calls :	Number of customer service calls made

2. COMPLETITUD

   Obtenga la completitud por variables y añada los resultados al PDF de resultados.

3. TRANSFORME LOS MISSINGS

   Valores fuera de la naturaleza de la variable o con un formato distinto deben ser reemplazados por NaN. Realice el conteo de estos y cree una tabla con los valores por variable. Genere nuevamente la tabla de completitud.

	columna	n_valores_re_nan
0	state	2
1	account length	4
2	area code	6
3	phone number	8
4	international plan	10
5	voice mail plan	12
6	number vmail messages	14
7	total day minutes	16
8	total day calls	18
9	tal day charge	20
10	total eve minutes	22
11	total eve calls	24
12	tal eve charge	26
13	total night minutes	28
14	total night calls	30
15	total night charge	32
16	total intl minutes	34
17	total intl calls	36
18	total intl charge	38
19	customer service calls	40

4. OUTLIERS

   Identifique los outliers por los siguientes tres métodos y realice una tabla de resultados , como se muestra al final.

   - IQR
   - Percentiles (Elementos que estén por debajo del percentil 5 y por encima del percentil 95)
   - Mean Change
   - Z-Score (Si es posible)

	features	n_outliers_IQR	n_outliers_Percentil	n_outliers_Z_Score	n_outliers_IQR_%	n_outliers_Percentil_%	n_outliers_Z_Score_%	total_outliers	%_outliers	indices
0	c_account length	18	323	7	0.54	9.69	0.21	18	0.54	[416, 1408, 197, 1093, 2150, 2277, 1387, 2700,...
1	c_number vmail messages	1	163	3	0.03	4.89	0.09	3	0.09	[2716, 845, 2887]
2	c_total day minutes	29	333	5	0.87	9.99	0.15	29	0.87	[15, 1679, 2836, 2971, 1052, 156, 2594,

Donde

- **features**: Nombre de las columnas continuas
- **n_outliers_IQR**: Número de outliers identificados a través de IQR
- **n_outliers_Percentil**: Número de outliers identificados a través de Percentiles
- **n_outliers_Z_Score**: Número de outliers identificados a través de ZScore
- **n_outliers_Mean_Change**: Número de outliers identificados a través de MeanChange
- **n_outliers_IQR_%** : Porcentaje de outliers por IQR respecto al total de registros
- **n_outliers_Percentil_%** : Porcentaje de outliers por Percentiles respecto al total de registros
- **n_outliers_Z_Score_%** : Porcentaje de outliers por Z_Score respecto al total de registros
- **n_outliers_Mean_Change_%** : Porcentaje de outliers por Mean_Change respecto al total de registros
- **total_outliers**: Corresponde al total de valores que fueron identificados como Outliers por al menos dos métodos
- **%_outliers** : Porcentaje de de valores que fueron identificados como Outliers por al menos dos métodos
- **indices**: Indices de la tabla que son considerados como Outliers por al menos dos métodos

Remueva los registros que sean identificados como Outliers por al menos dos métodos variable por variable. Muestre un cuadro con el número de registros antes y después de la remoción de los mismos variable por variable.

	v_feature	c_n_rows
0	Inicial	3333
1	c_account length	3315
2	c_number vmail messages	3312
3	c_total day minutes	3277
4	c_total day calls	3241
5	c_total eve minutes	3197
6	c_total eve calls	3177
7	c_total night minutes	3141
8	c_total night calls	3122
9	c_total night minutes	3122
10	c_total night calls	3122
11	c_total night charge	3122
12	c_total intl minutes	3112
13	c_total intl calls	3037
14	c_total intl charge	2999
15	c_customer service calls	2979

5. Genere los histogramas de las variables continuas antes y despues de cada tratamiento de Outliers, añada los gráficos al PDF de resultados.

6. TRATAMIENTO DE VALORES AUSENTES (En el PDF debe haber una descripción de esta sección)

   - Elimine aquellas columnas que superen el umbral de 20% o mas de presencia de valores ausentes. Indique qué columnas fueron eliminadas.
   - Genere su conjunto de entrenamiento y test , donde el test debe poseer el 20% de la información. (La moda, mediana, media deben obtenerse del conjunto de entrenamiento e imputar tanto en train como en test)
   - Impute las variables discretas que lo requieran mediante el uso de la moda, realice el test chi-cuadrado solo para confirmar que las frecuencias se mantienen. La variable imputada debe almacenarse en una variable nueva, dejando a la original intacta.
   - Impute las variables continuas que lo requieran mediante el uso de la mediana, moda o media, seleccione la mejor opción. La variable imputada debe almacenarse en una variable nueva, dejando a la original intacta.

ENTREGABLES:
-PDF con tablas, gráficos y pequeñas descripciones
-Código en Python (ordenado, limpio y sin errores)

PARTE 2. REDUCCIÓN DE DIMENSIONALIDAD
CONJUNTO DE DATOS
Tabla que posee información que es el resultado de un análisis químico de vinos cultivados en la misma región en Italia por tres cultivadores diferentes. Hay trece medidas diferentes tomadas para diferentes componentes que se encuentran en los tres tipos de vino. El objetivo es conocer el cultivador a partir de las características del vino.

Tabla "WINE"

state	account length	area code	phone number	internati onal plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls	churn
KS	128	415	382-4657	no	yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	3	2.7	1	FALSE
OH	107	415	371-7191	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7	1	FALSE
NJ	137	415	358-1921	no	no	0	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29	0	FALSE
OH	84	408	375-9999	yes	no	0	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	FALSE
OK	75	415	330-6626	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	FALSE
AL	118	510	391-8027	yes	no	0	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.7	0	FALSE
MA	121	510	355-9993	no	yes	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	02.03	3	FALSE
MO	147	415	329-9001	yes	no	0	157	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92	0	FALSE
LA	117	408	335-4719	no	no	0	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35	1	FALSE
WV	141	415	330-8173	yes	yes	37	258.6	84	43.96	222	111	18.87	326.4	97	14.69	11.2	5	03.02	0	FALSE
IN	65	415	329-6603	no	no	0	129.1	137	21.95	228.5	83	19.42	208.8	111	9.4	12.7	6	3.43	4	TRUE
RI	74	415	344-9403	no	no	0	187.7	127	31.91	163.4	148	13.89	196	94	8.82	9.1	5	2.46	0	FALSE

Parte 2.1

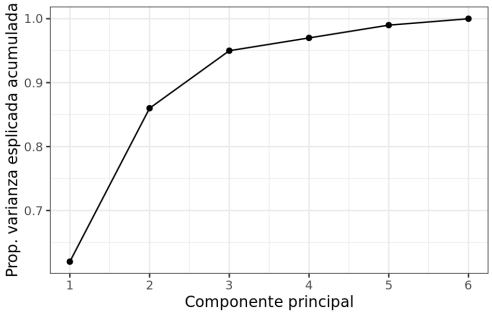
- Utilizando "Valor perdido" , ¿Qué variables se eliminan de acuerdo a este método? Elimine esa variable de forma definitiva para no tener problemas el los siguientes puntos.
- Utilizando "Baja Varianza" , ¿Qué variables se eliminan de acuerdo a este método? (varianza inferior a .01)
- Utilizando "Alta Correlación" , ¿Qué variables se eliminan de acuerdo a este método (Define el umbral para que se considere como correlación alta)? Añada mapas de calor.

- Utilizando "Correlación (con el objetivo)" , ¿Qué variables se eliminan de acuerdo a este método (Si la correlación entre una variable y la tgt es menor a 0.1)? Añada mapas de calor.

- Utilizando "Multicolinealidad" , ¿Qué variables se eliminan de acuerdo a este método?, agregue el VIF asociado a esas variables.

Parte 2.2

- Genere su conjunto de entrenamiento y prueba, el conjunto de prueba debe corresponder al 30% de los datos
- Utilice PCA para reducción de dimensiones , pruebe con distintos números de componentes, desde 2 hasta 10 y realice un gráfico como el que se muestra a continuación.



- ¿Cuántos componentes son suficientes de tal forma que los componentes representan el 80% de la varianza?

- Adicionalmente muestre como se ve el conjunto de datos en tres dimensiones , haciendo una distinción entre el cultivador (tgt)

- Utiliza SelectKBest , Seleccione las mejores 10 variables

ENTREGABLES: -PDF con resultados -Notebook ordenado , con un distintivo por punto requerido.