

Reporte de Congestión Vial

Autores: Equipo de Analítica

Curso: Minería de Datos

Fecha: 21/09/2025

Introducción

Este reporte resume el flujo de análisis aplicado sobre el dataset de congestión vehicular de Santiago del 14/03/2025.

Se desarrollaron etapas de limpieza, exploración, reducción de dimensionalidad y modelado predictivo con enfoque reproducible en Python.

Metodología

La metodología incluyó depuración de datos temporales, creación de variables de duración y períodos peak, evaluación de normalidad, PCA y modelado lineal.

El conjunto analizado contiene 3520 registros y 21 variables tras el enriquecimiento.

Análisis Exploratorio (EDA)

Se generaron estadísticas descriptivas para variables numéricas y categóricas, gráficos de distribución, correlaciones y análisis por comuna.

Los resultados completos se presentan en las tablas y figuras asociadas a esta sección.

Evaluación de Normalidad

Prueba de normalidad: Shapiro-Wilk Variable: Velocidad km/h Estadístico: 0.9715 p-valor: 0.0000 n: 3520 Conclusión: Se rechaza la normalidad al 5%.

Análisis PCA

Se requieren 6 componentes (hasta PC6) para explicar al menos el 95% de la varianza.

Se evaluaron las cargas de las dos primeras componentes para interpretar patrones espaciales y temporales.

Modelo 1

Modelo lineal con todas las variables disponibles (preprocesamiento estándar y codificación one-hot).

Desempeño: RMSE: 4.082, R2: 0.151, MAE: 3.339, Bias: -0.179.

Modelo 2

Modelo lineal con selección secuencial de variables para mejorar interpretabilidad.

Desempeño: RMSE: 4.089, R2: 0.148, MAE: 3.332, Bias: -0.174.

Comparación de Modelos

Se compararon métricas clave frente a un baseline simple y se generaron gráficos diagnósticos para ambos modelos.

tabla resumen numericas

	variable	count	mean	std	min	q25	median	q75	max	skew	kurtosis
	X	3520.0	-70.66	0.11	-71.47	-70.71	-70.64	-70.58	-70.5	-2.48	9.44
	Y	3520.0	-33.47	0.1	-33.91	-33.53	-33.45	-33.42	-33.0	-0.53	1.8
	Z	3520.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ranking Regional		3520.0	6485.47	3388.48	4.0	3681.0	6466.5	9174.5	13459.5	0.03	-1.0
	Largo km	3520.0	0.39	0.28	0.11	0.23	0.32	0.46	8.29	8.23	177.65
Velocidad km/h		3520.0	17.89	4.43	0.53	14.79	18.2	21.49	25.0	-0.44	-0.4
duracion_horas		3520.0	0.83	1.0	0.25	0.33	0.5	0.92	9.25	3.86	19.3
	duracion_min	3520.0	49.79	59.97	15.0	19.8	30.0	55.2	555.0	3.86	19.3
hora_inicio_num		3520.0	13.12	4.29	6.17	8.08	13.46	16.83	20.92	-0.03	-1.39
	hora_fin_num	3520.0	13.95	4.41	6.5	8.83	14.25	18.0	23.42	-0.05	-1.42

tabla categoricas

variable	n_distinct	top	freq_top
Calle	1348	Mariano Sánchez Fontecilla	25
Comuna	48	Las Condes	297
ID	3520	Maipú-692	1
peak	2	Punta	1862

tabla faltantes

variable	n_missing	pct_missing
X	0	0.0
Y	0	0.0
Z	0	0.0
Calle	39	1.11
Comuna	0	0.0
Fecha	0	0.0
Hora Fin	0	0.0
Hora Inicio	0	0.0
ID	0	0.0
Ranking Regional	0	0.0
n	0	0.0
Largo km	0	0.0
Velocidad km/h	0	0.0
dt_inicio	0	0.0
dt_fin	0	0.0
duracion_horas	0	0.0
duracion_min	0	0.0
hora_inicio_num	0	0.0
hora_fin_num	0	0.0
peak	0	0.0
TARGET_BIN	0	0.0

tabla grupo comuna

Comuna	n_tramos	vel_prom	vel_std	largo_prom	duracion_min_prom
Buin	48	18.86	3.85	0.46	65.64
Calera de Tango	11	18.29	3.95	0.48	51.82
Cerrillos	45	17.95	3.48	0.4	50.31
Cerro Navia	20	20.09	4.04	0.44	55.5
Colina	91	18.21	3.96	0.4	39.95
Conchalí	42	19.09	3.78	0.37	42.84
Curacaví	4	16.03	4.26	0.35	22.5
El Bosque	32	18.07	3.8	0.43	54.24
El Monte	7	17.22	3.05	0.38	43.63
Estación Central	65	19.06	3.92	0.41	53.88
Huechuraba	56	17.21	4.69	0.41	44.53
Independencia	48	17.65	4.72	0.39	51.69
Isla de Maipo	1	18.33	nan	0.33	34.8
La Cisterna	72	18.29	3.52	0.34	39.83
La Florida	195	17.93	3.89	0.36	37.55
La Granja	43	17.03	3.79	0.32	45.84
La Pintana	43	18.62	4.19	0.35	41.64
La Reina	112	17.42	4.94	0.4	57.8
Lampa	43	17.31	5.4	0.56	54.53
Las Condes	297	18.03	4.14	0.39	49.42
Lo Barnechea	43	18.27	4.6	0.4	34.4
Lo Espejo	24	19.24	3.52	0.28	48.15
Lo Prado	27	17.9	4.86	0.37	62.4
Macul	51	17.59	3.7	0.33	36.35
Maipú	194	17.64	4.54	0.42	51.45
Melipilla	65	19.02	4.5	0.52	64.55
Padre Hurtado	27	16.75	4.35	0.35	40.47
Paine	32	18.63	4.21	0.43	29.19
Pedro Aguirre Cerda	25	20.46	3.57	0.36	39.0
Peñaflor	34	18.21	4.33	0.37	36.88
Peñalolén	139	16.78	4.12	0.38	60.61
Pirque	12	18.5	4.63	0.87	47.9
Providencia	246	17.24	5.09	0.35	56.11
Pudahuel	103	15.37	5.71	0.41	59.12
Puente Alto	173	17.83	4.61	0.36	46.64
Quilicura	103	17.57	4.26	0.37	50.25
Quinta Normal	42	16.84	5.04	0.38	57.27
Recoleta	97	16.93	5.41	0.41	67.92
Renca	56	18.05	4.42	0.39	47.91
San Bernardo	115	18.11	4.57	0.44	61.02
San Joaquín	20	20.18	3.57	0.39	44.73
San Miguel	65	18.35	3.28	0.36	47.2
San Pedro	2	19.26	1.64	0.3	15.0
San Ramón	39	17.15	3.81	0.31	42.82
Santiago	194	18.91	3.99	0.46	51.59
Talagante	32	18.52	3.73	0.4	43.91
Vitacura	141	18.03	4.19	0.35	43.45
Ñuñoa	144	18.56	4.57	0.39	41.7

correlacion top pairs

var1	var2	corr
duracion_horas	duracion_min	0.9999999999999968
hora_inicio_num	hora_fin_num	0.9739355332682422
Ranking Regional	Largo km	-0.6897624194106322
Ranking Regional	duracion_min	-0.5406030400573674
Ranking Regional	duracion_horas	-0.5406030400573661

tabla pca varianza

componente	var_exp	var_exp_acum
PC1	31.92	31.92
PC2	22.47	54.39
PC3	15.17	69.56
PC4	13.31	82.87
PC5	7.39	90.26
PC6	6.78	97.03
PC7	2.97	100.0
PC8	0.0	100.0
PC9	0.0	100.0
PC10	0.0	100.0

tabla pca cargas pc1

variable	loading	abs_loading_rank
duracion_horas	0.5193068003025956	1
duracion_min	0.5193068003025955	2
Ranking Regional	-0.4868574927824867	3
Largo km	0.4094586860570753	4
Velocidad km/h	0.1756003740856113	5
hora_fin_num	0.1432092906842561	6
X	-0.0557550239220621	7
Y	-0.0286106858869257	8
hora_inicio_num	0.0261280327648269	9
Z	0.0	10

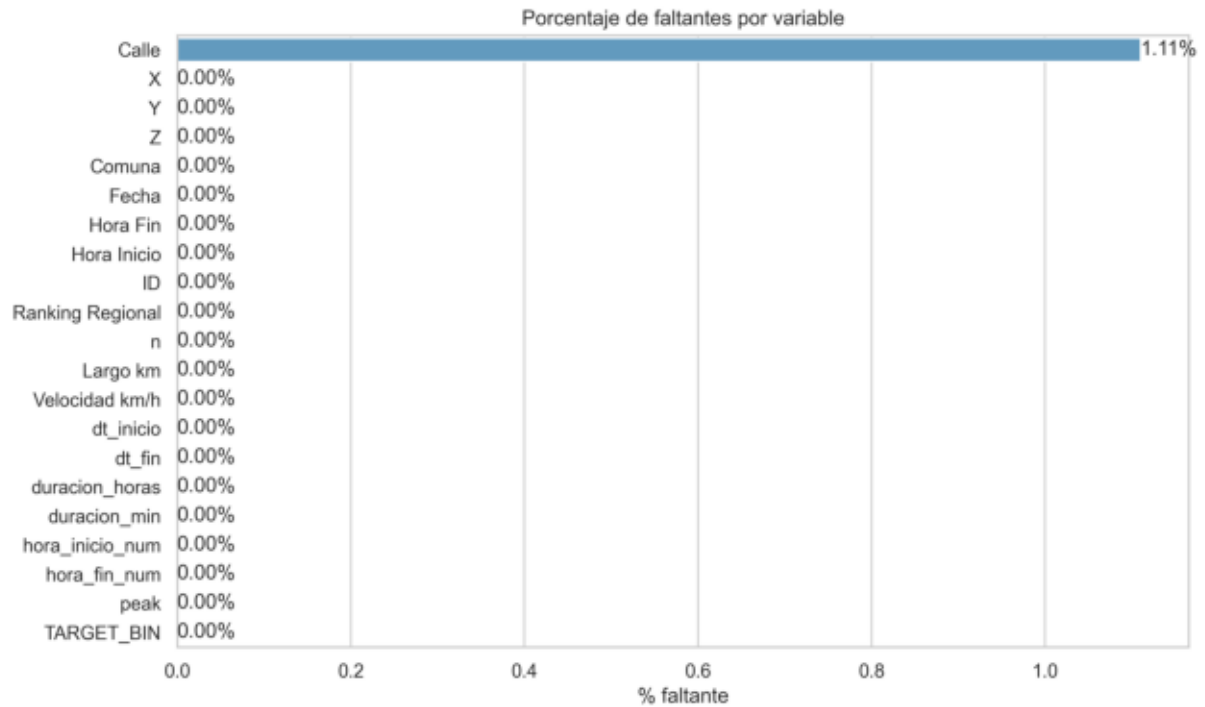
tabla pca cargas pc2

variable	loading	abs_loading_rank
hora_inicio_num	0.6828665947684414	1
hora_fin_num	0.6762725431888905	2
Largo km	-0.1800051124489062	3
Ranking Regional	0.1615698928527765	4
Velocidad km/h	-0.0951449160087308	5
Y	-0.0562644677171177	6
duracion_horas	0.0517048066114992	7
duracion_min	0.0517048066114992	8
X	0.0166424418496392	9
Z	6.938893903907227e-17	10

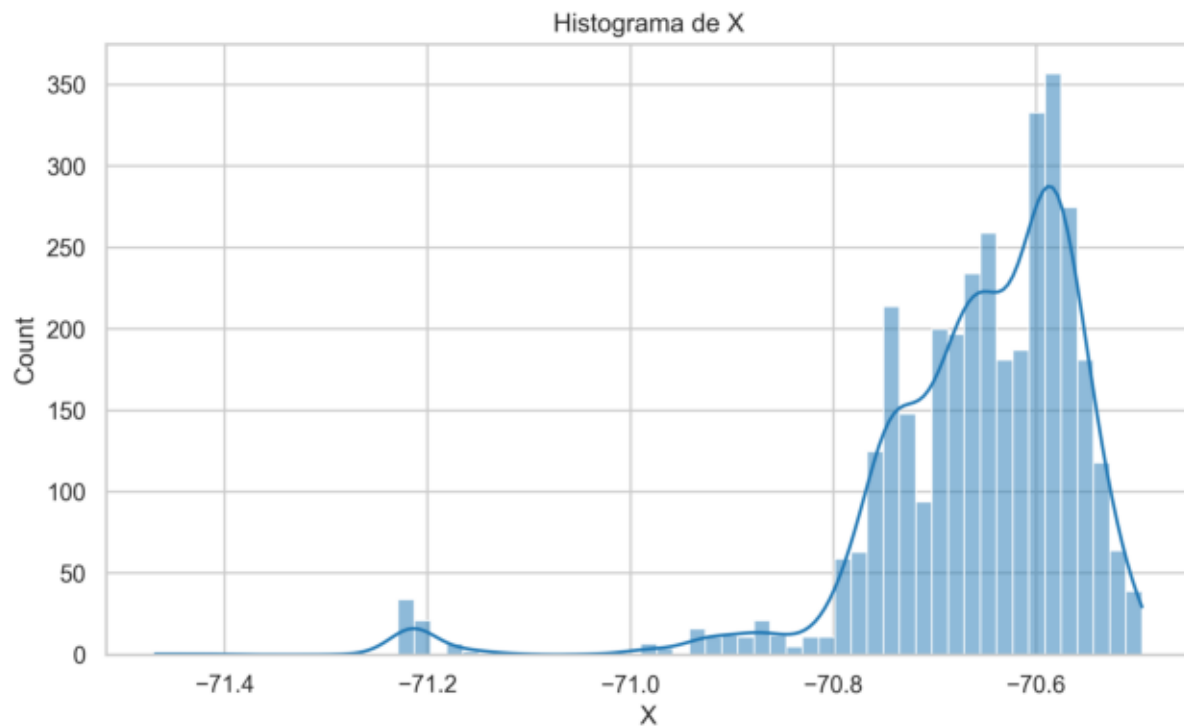
metricas modelos

modelo	RMSE	R2	MAE	Bias
baseline	4.432930176090961	-0.0008677215865517	3.687786232869936	-0.1305247138805066
modelo1	4.082424004788518	0.1511496737918791	3.338688453207356	-0.1788875684882789
modelo2	4.089393630805845	0.1482488386828758	3.3316577897639137	-0.1740132392777829

faltantes_bar.png

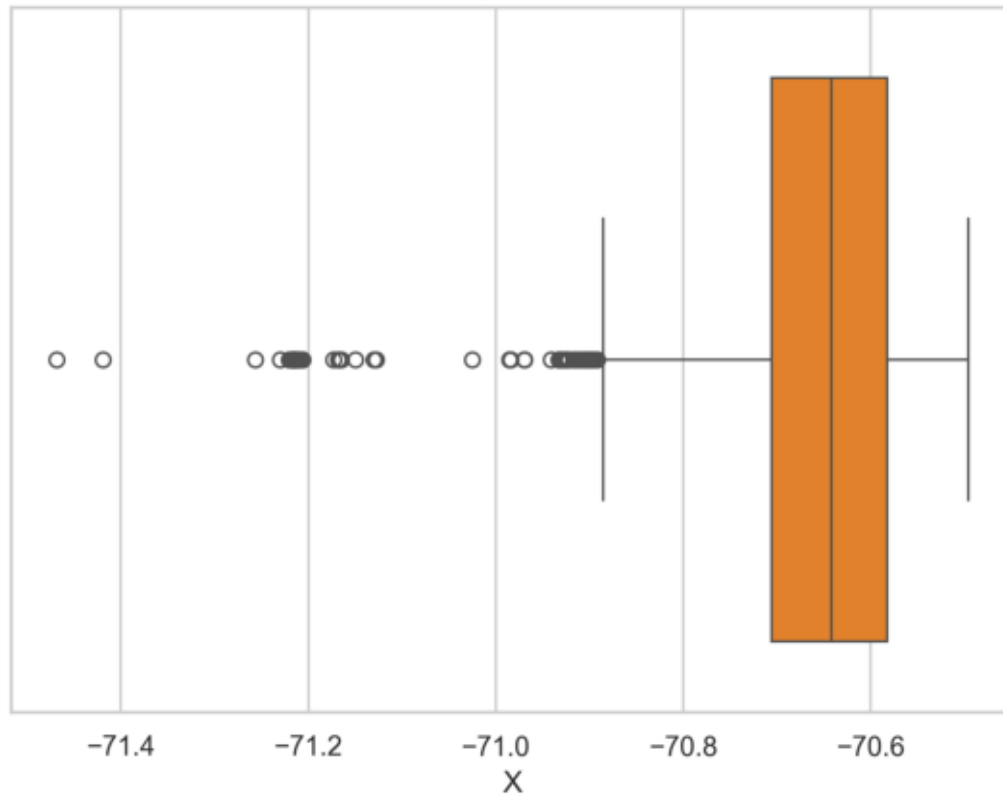


hist_X.png

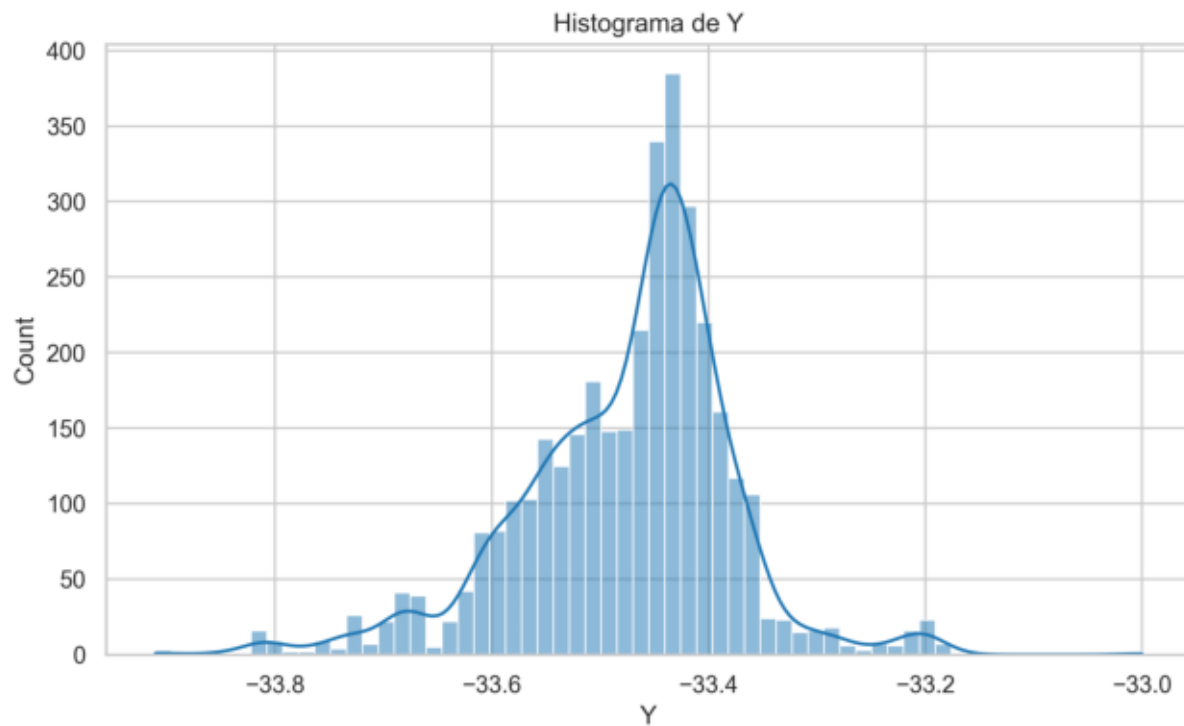


boxplot_X.png

Boxplot de X

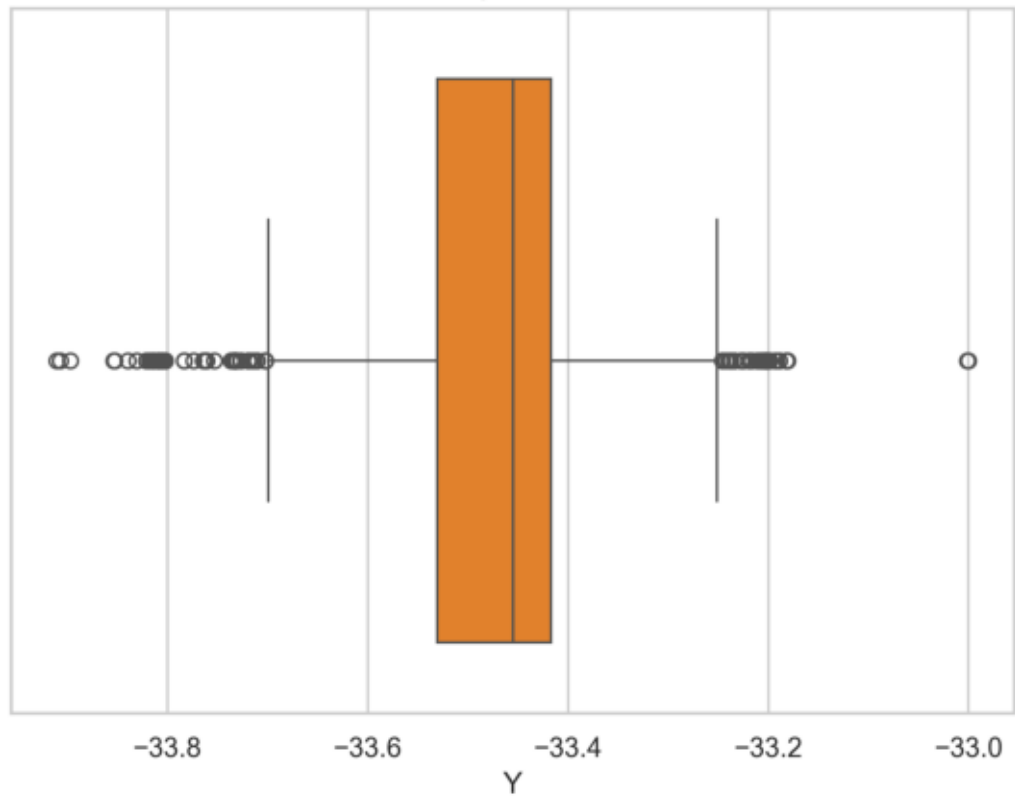


hist_Y.png

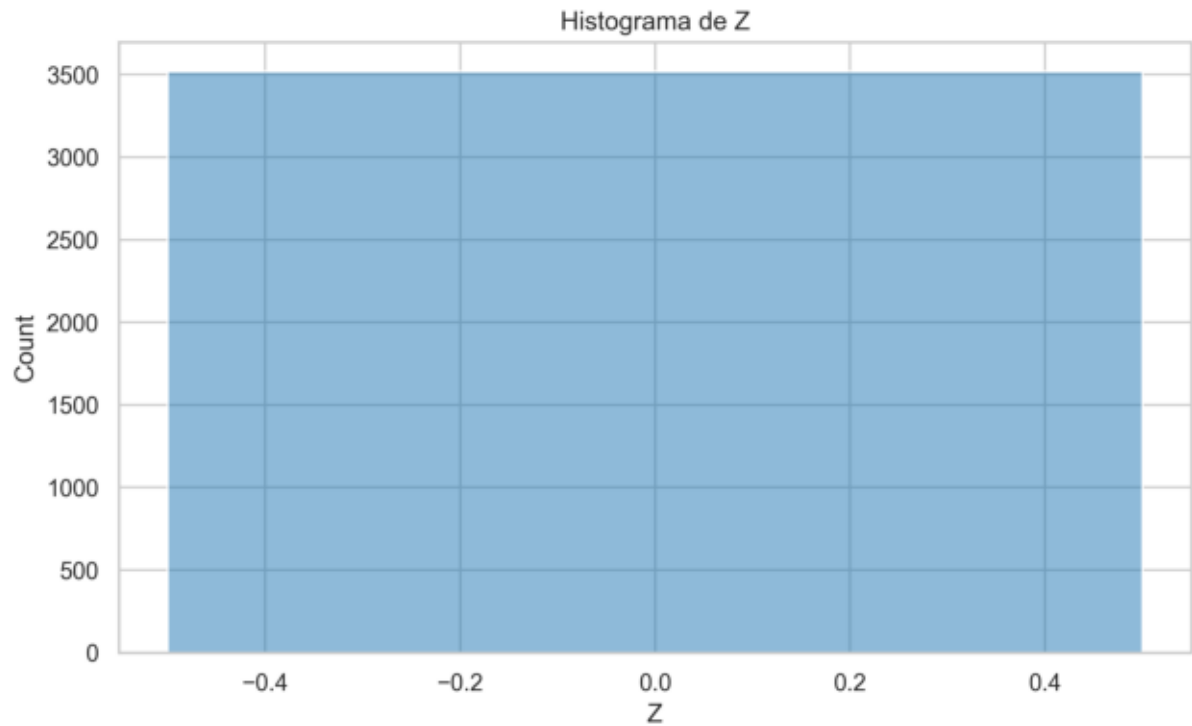


boxplot_Y.png

Boxplot de Y

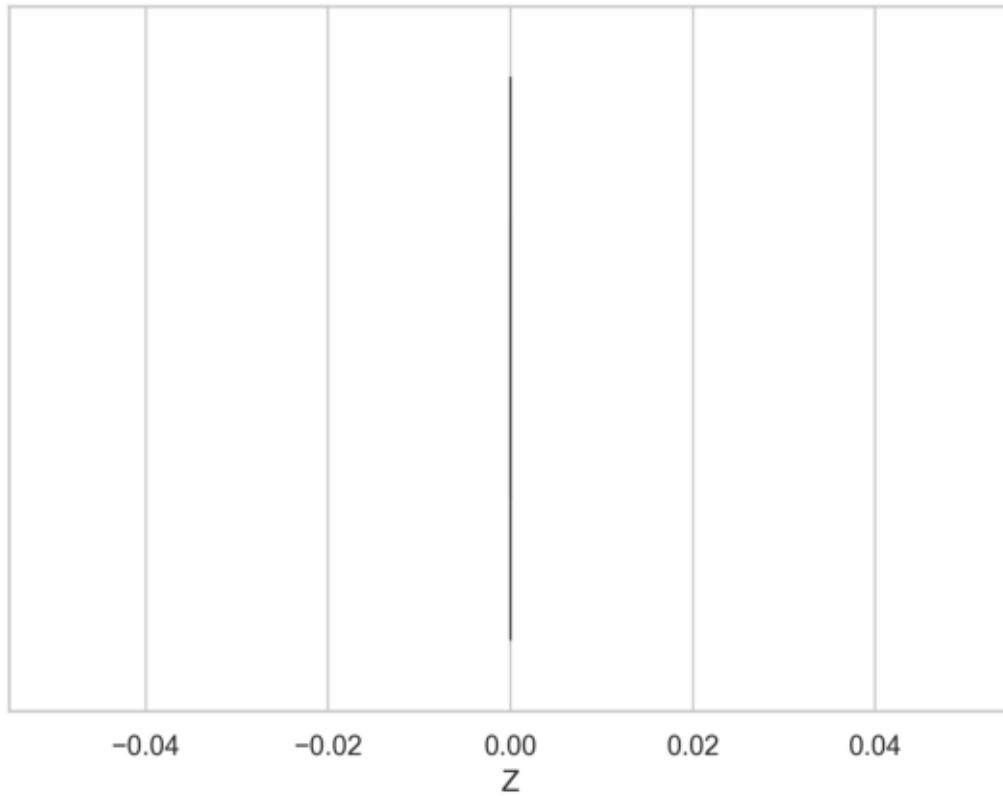


hist_Z.png

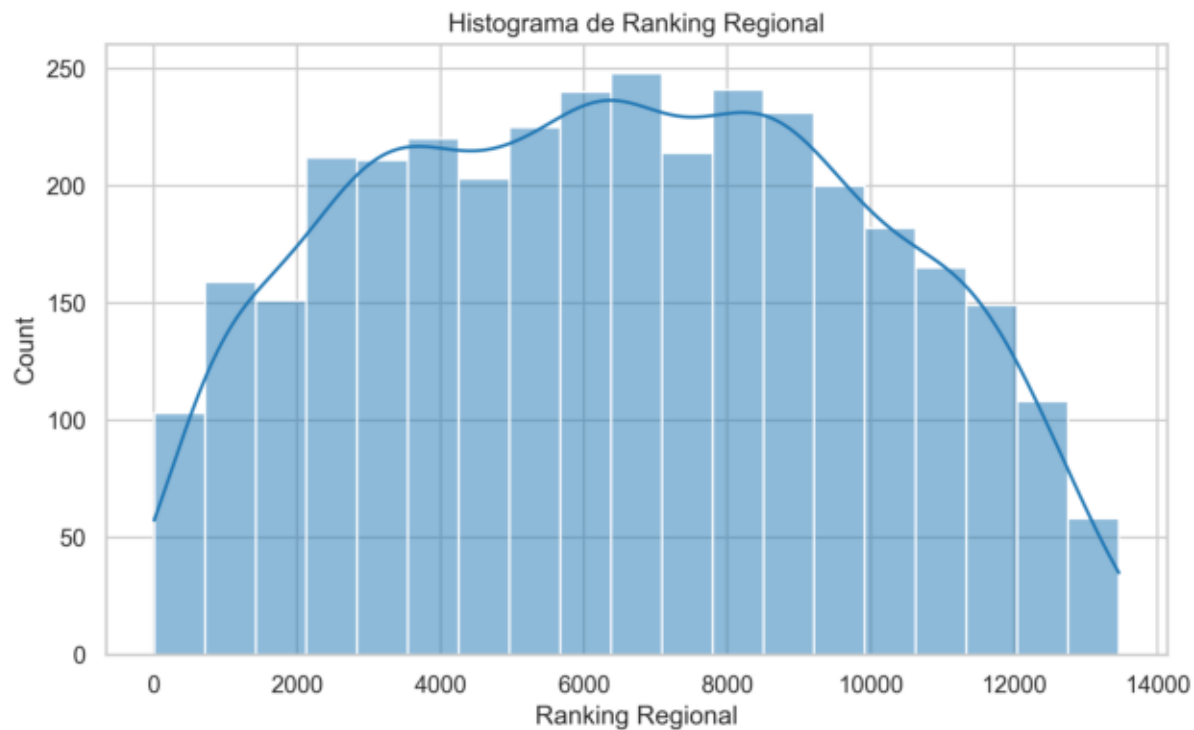


boxplot_Z.png

Boxplot de Z

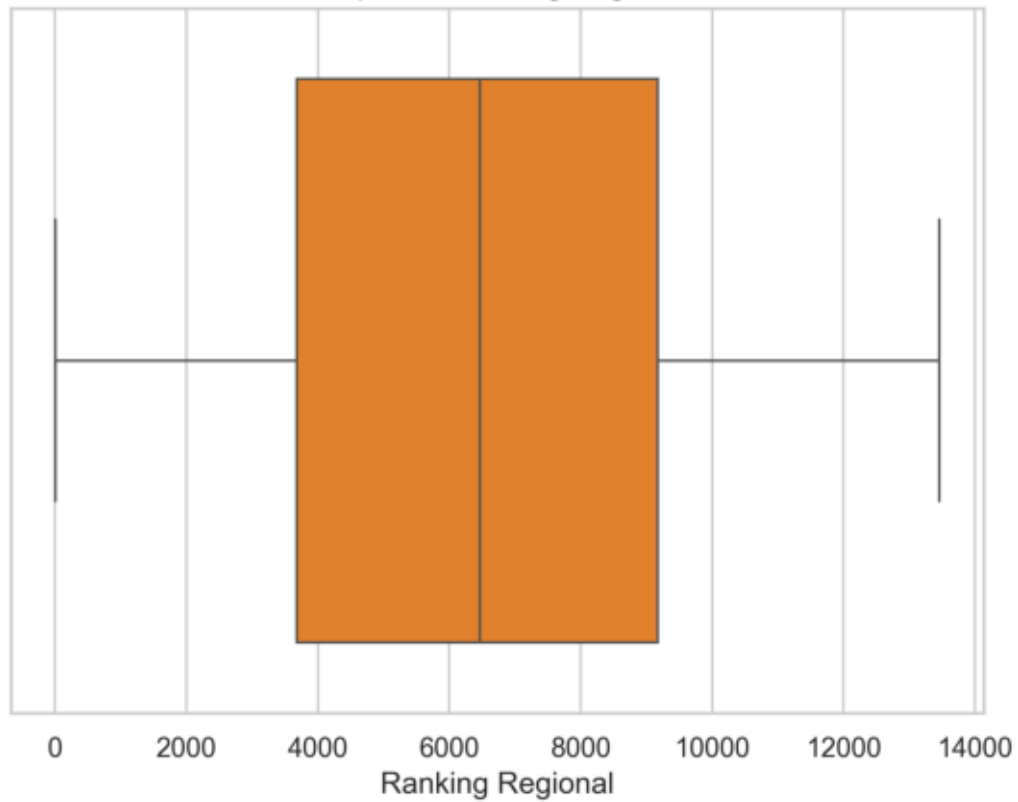


hist_Ranking_Regional.png

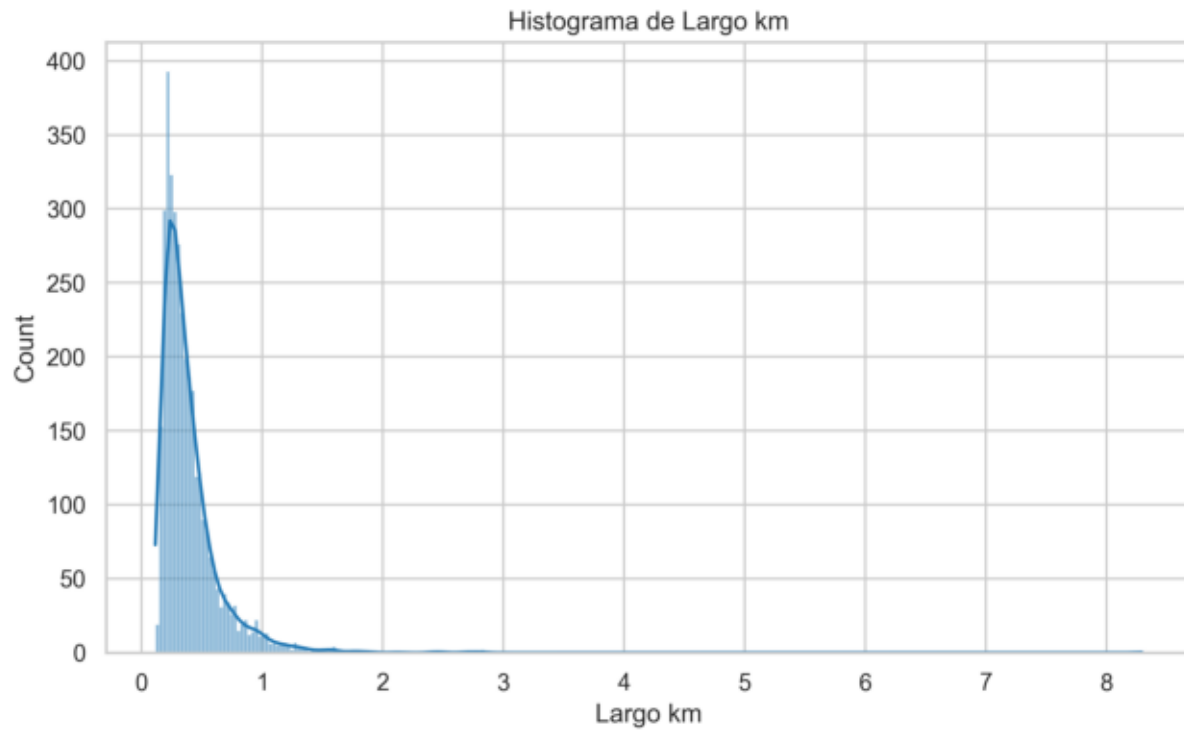


boxplot_Ranking_Regional.png

Boxplot de Ranking Regional

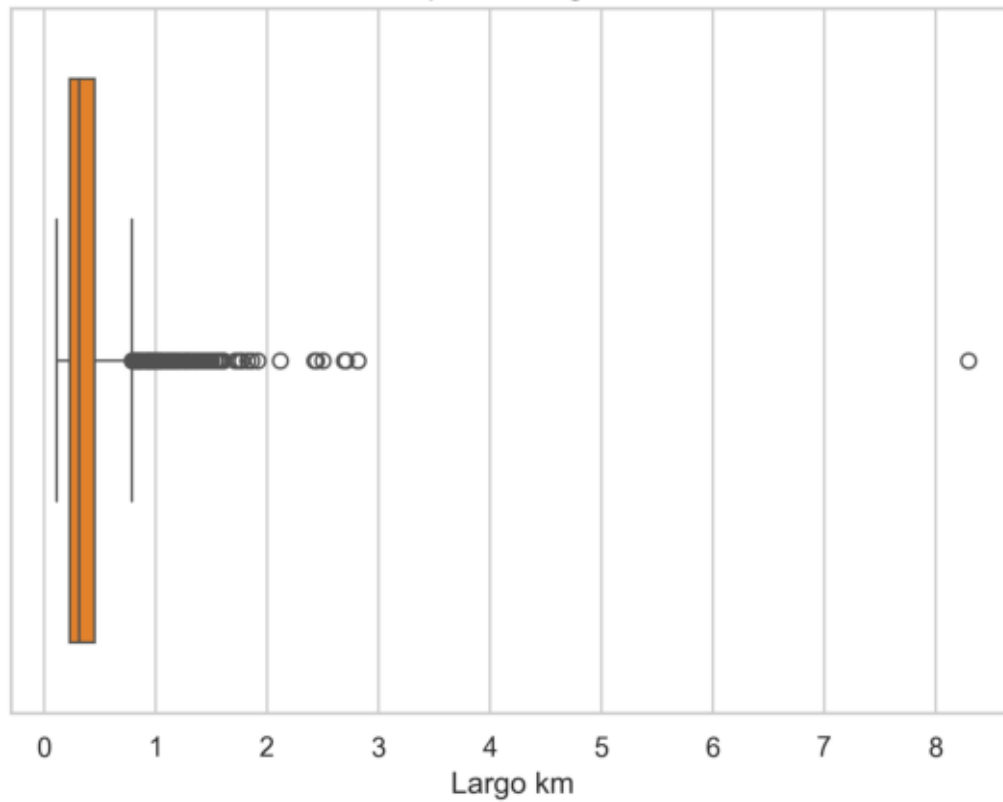


hist_Largo_km.png

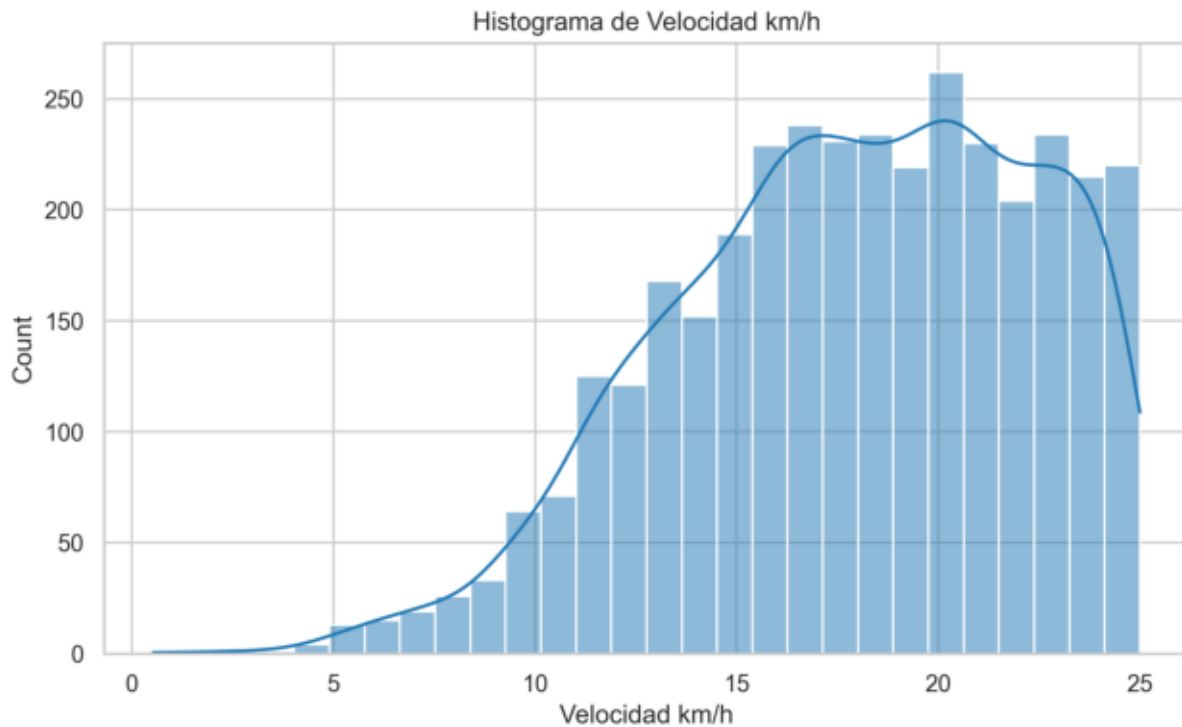


boxplot_Largo_km.png

Boxplot de Largo km

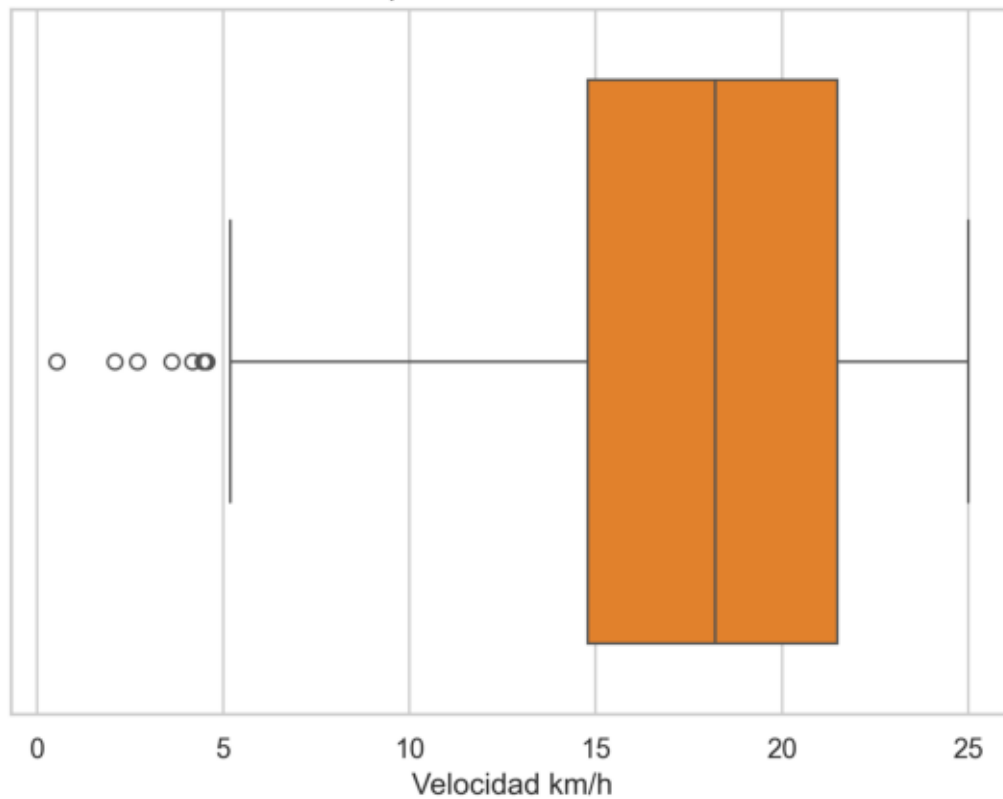


hist_Velocidad_kmh.png

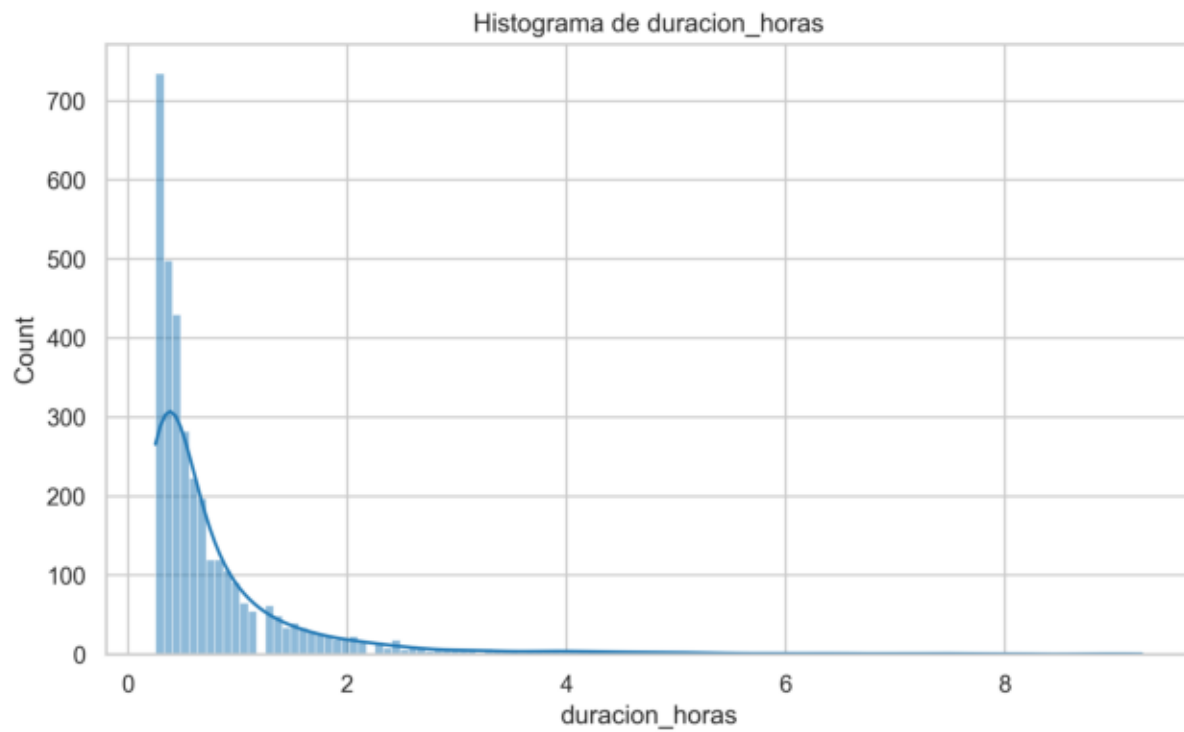


boxplot_Velocidad_kmh.png

Boxplot de Velocidad km/h

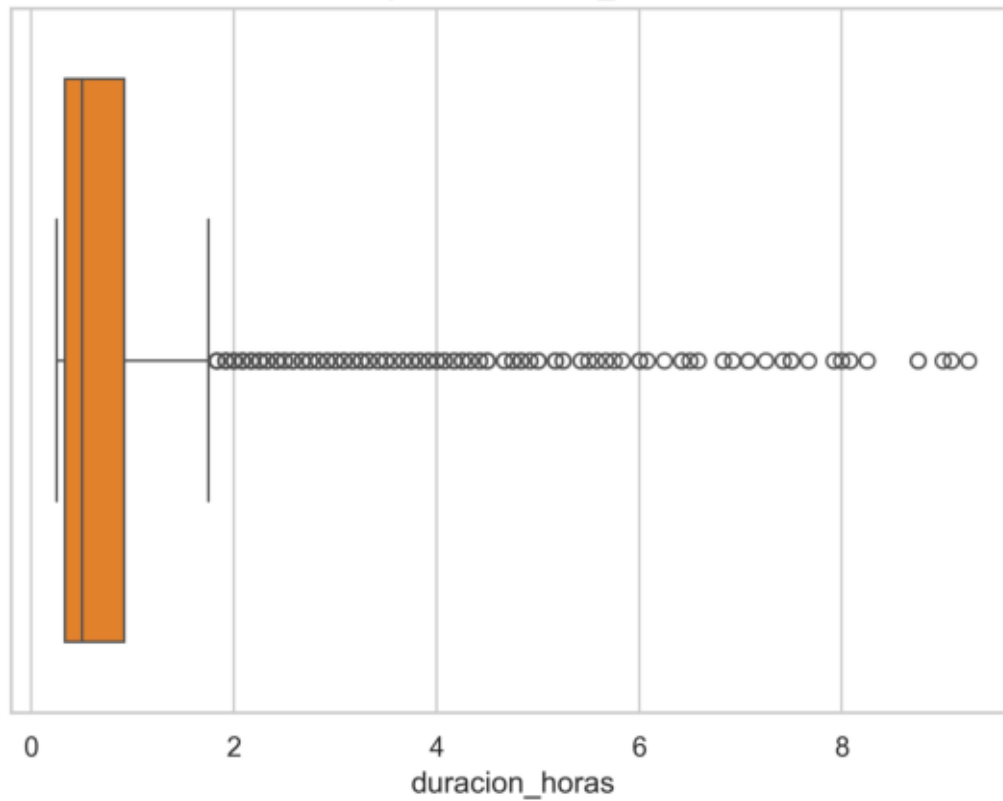


hist_duracion_horas.png

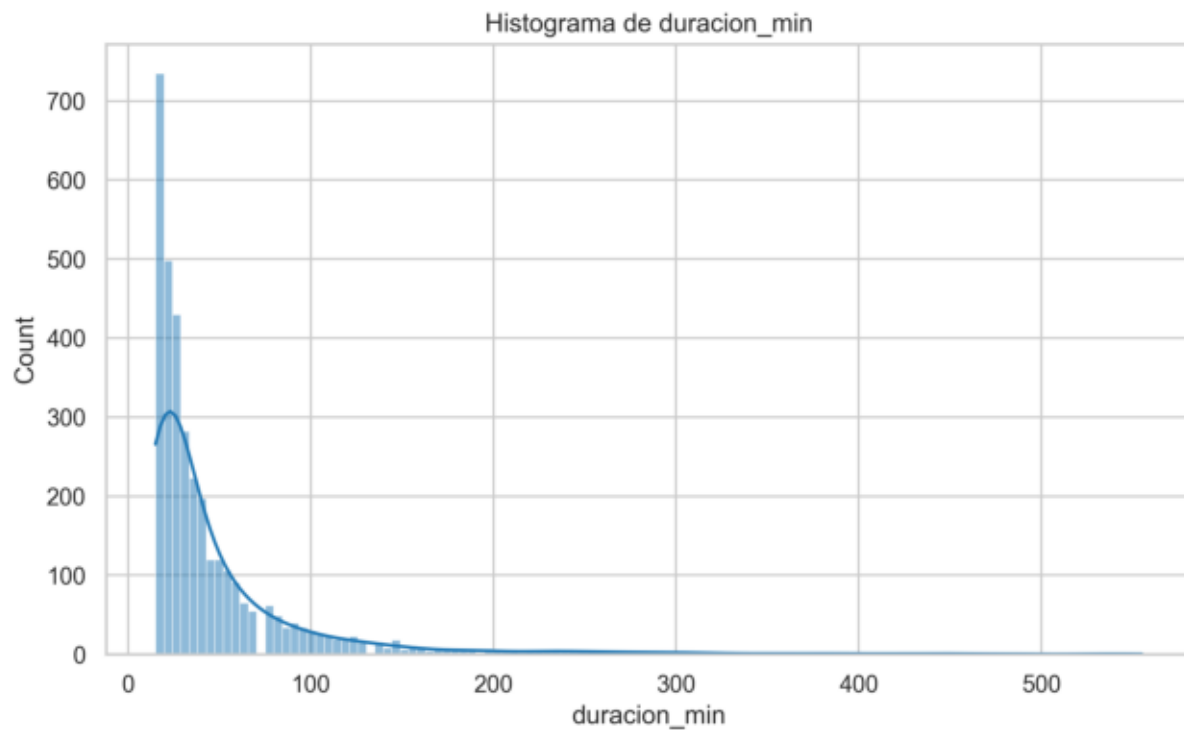


boxplot_duracion_horas.png

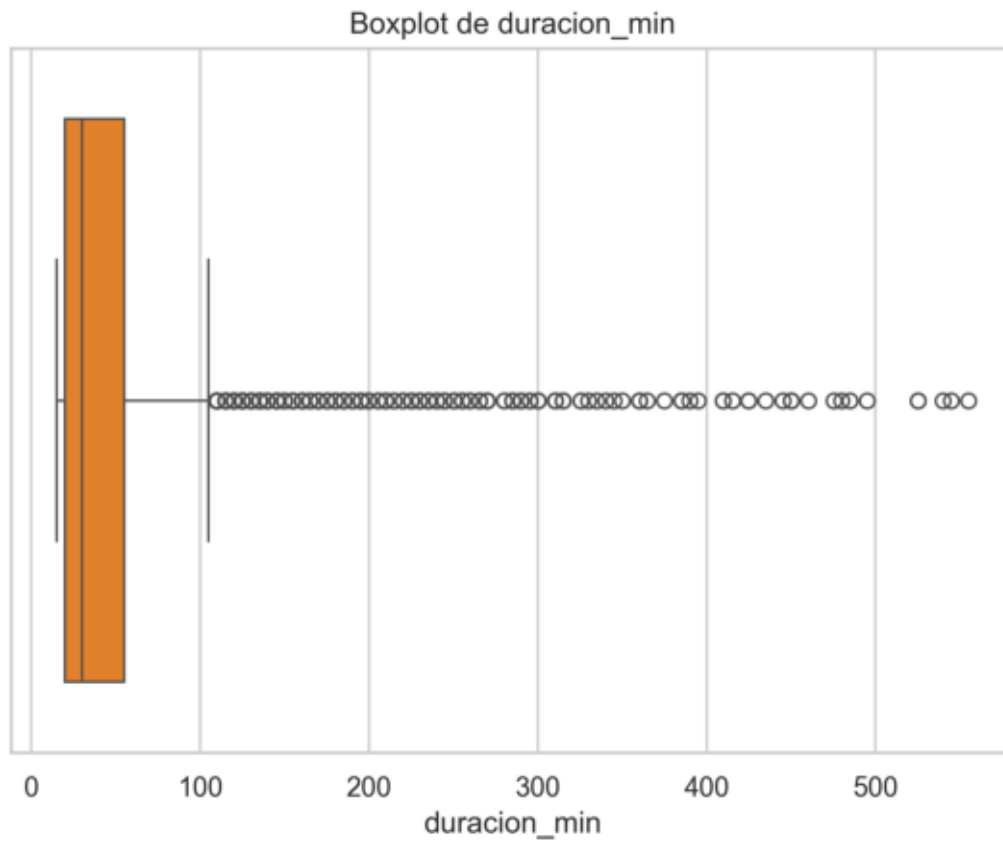
Boxplot de duracion_horas



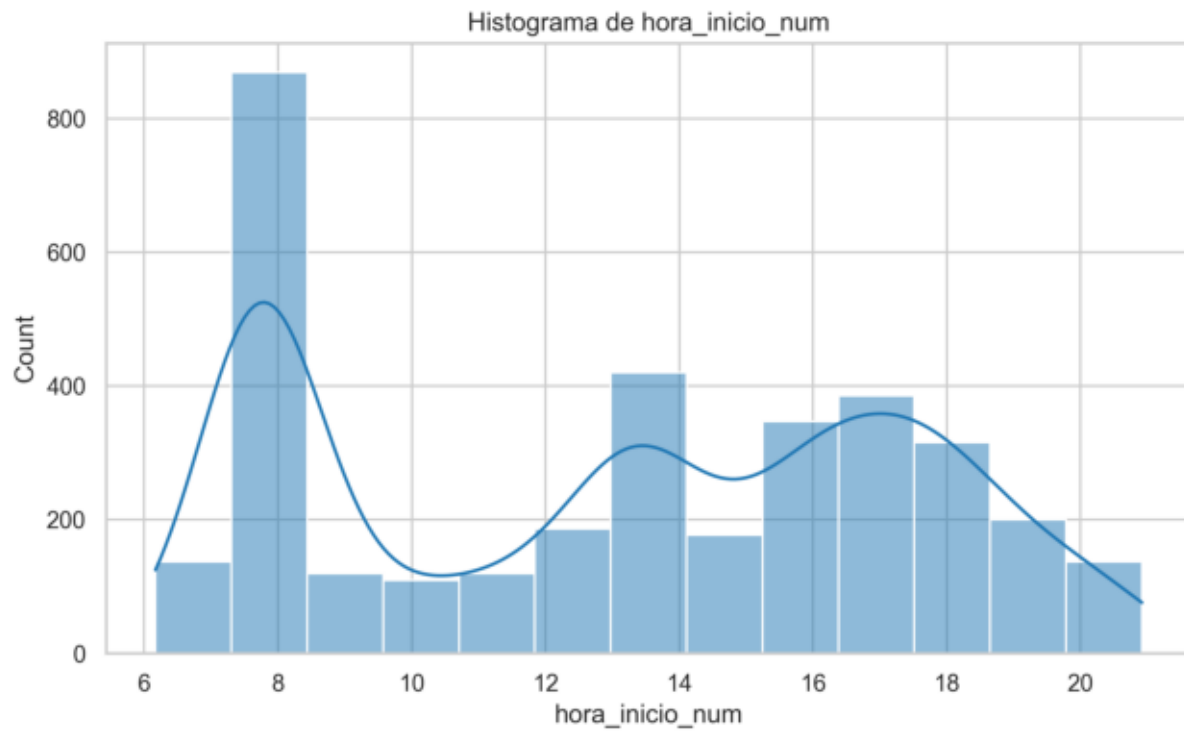
hist_duracion_min.png



boxplot_duracion_min.png

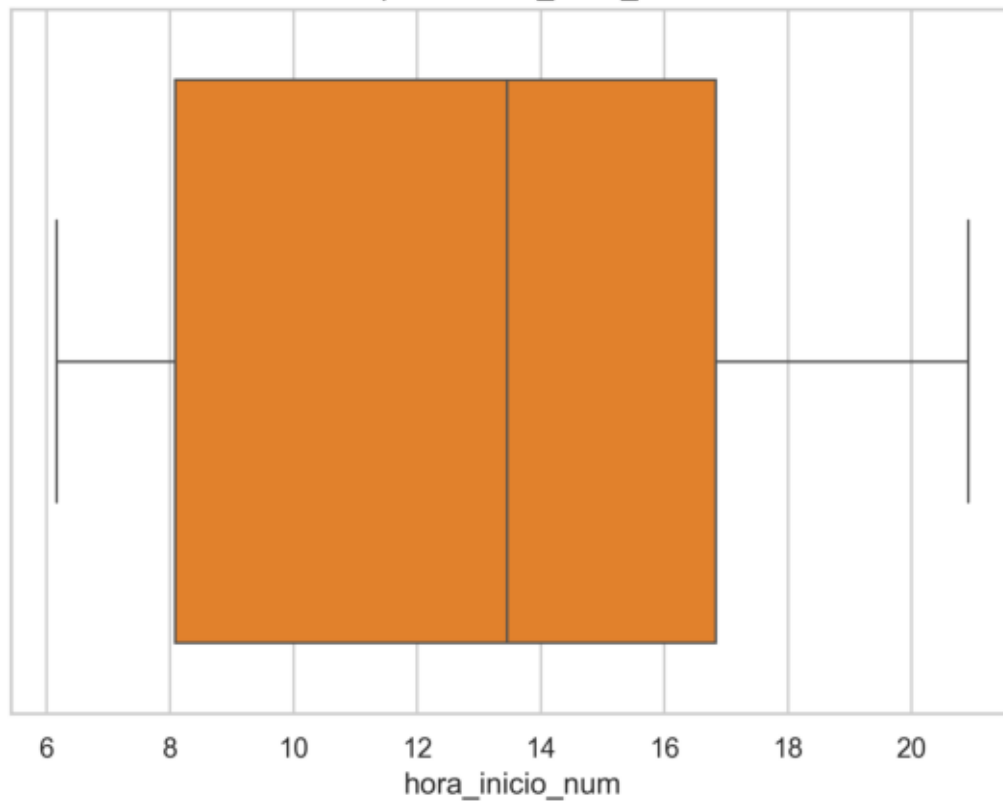


hist_hora_inicio_num.png

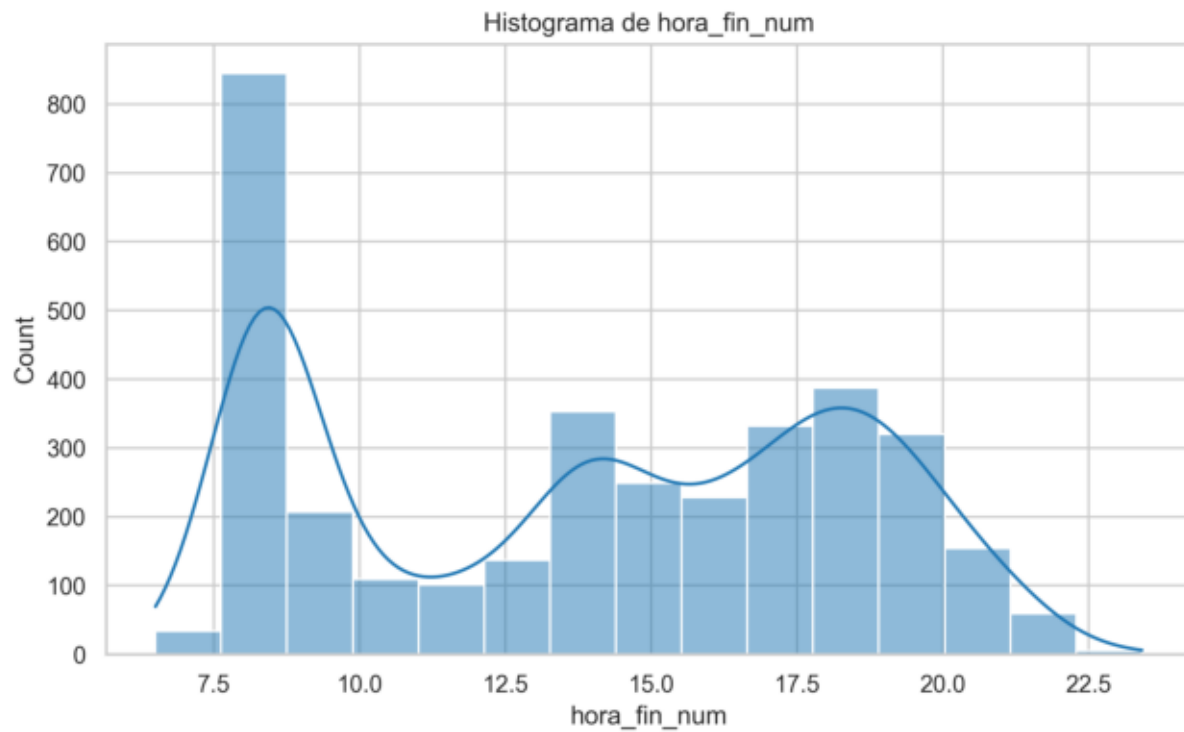


boxplot_hora_inicio_num.png

Boxplot de hora_inicio_num

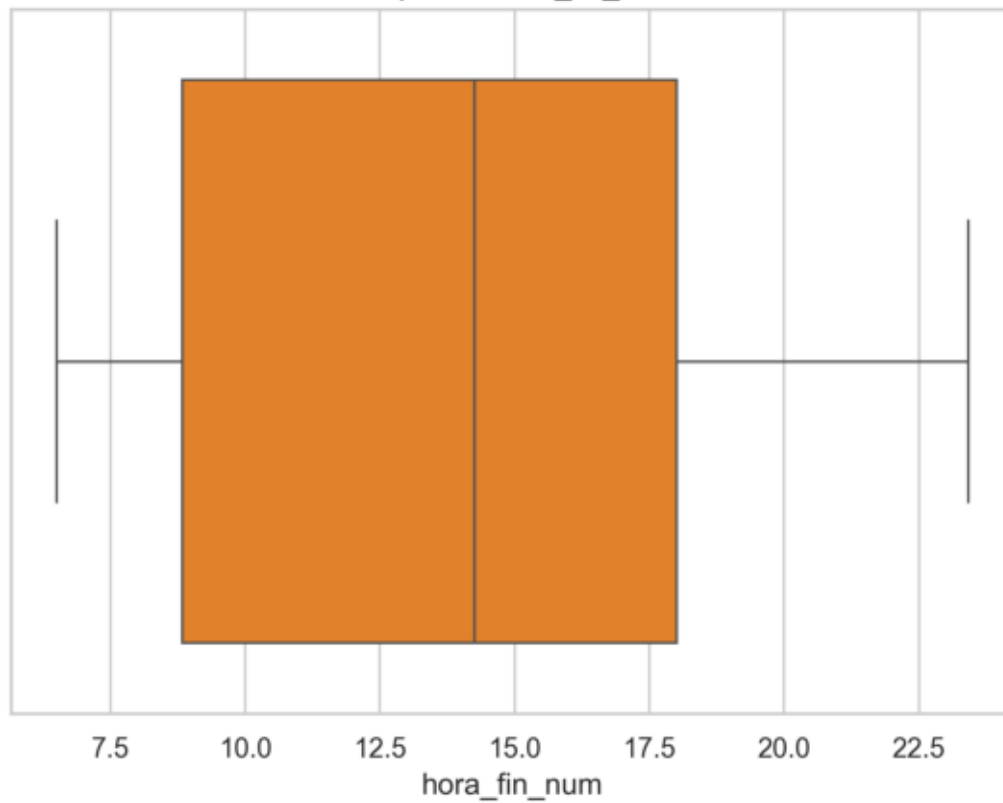


hist_hora_fin_num.png

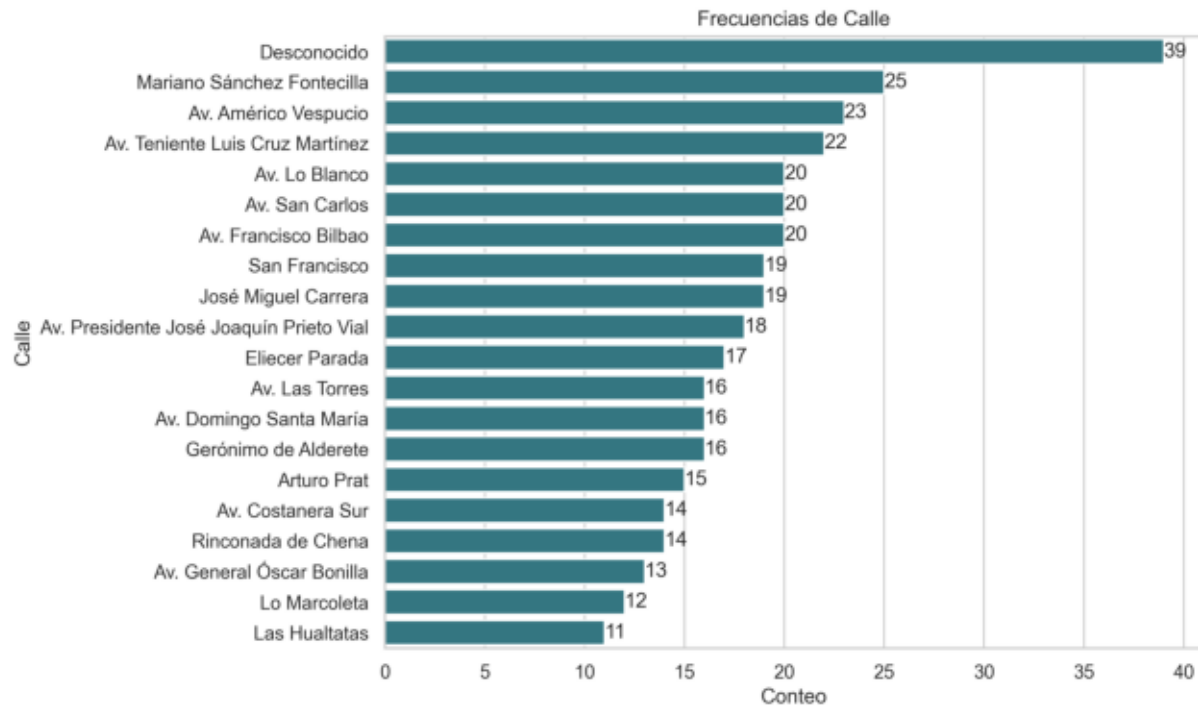


boxplot_hora_fin_num.png

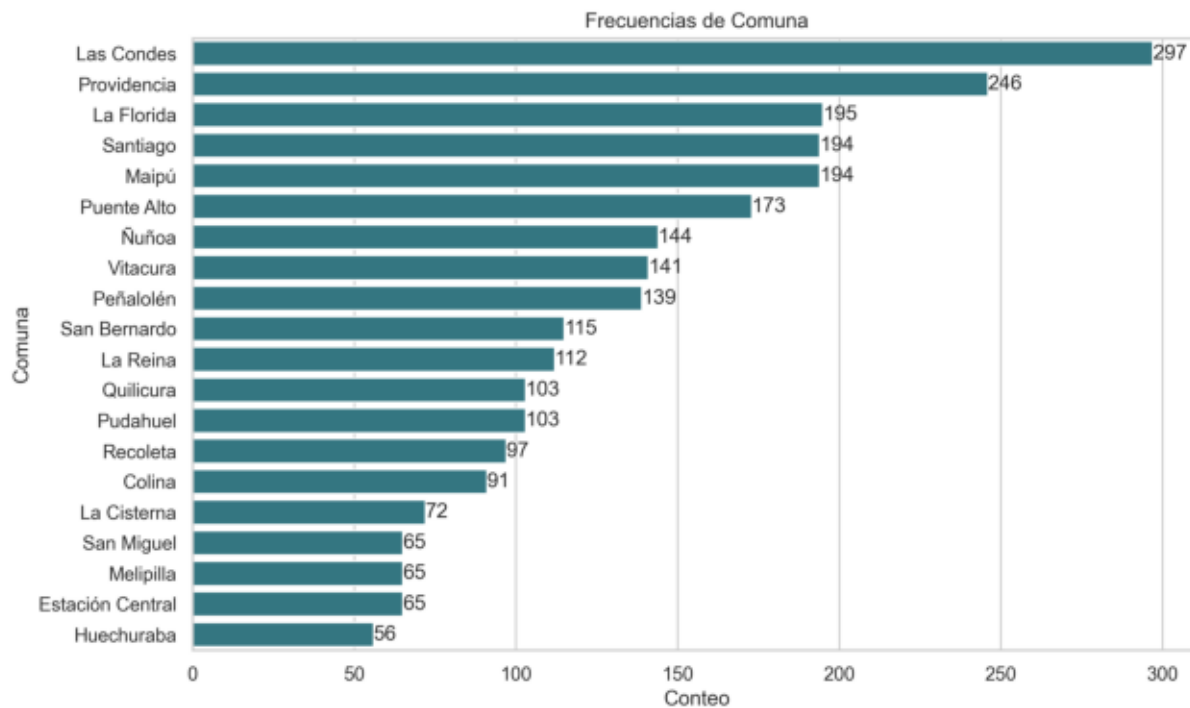
Boxplot de hora_fin_num



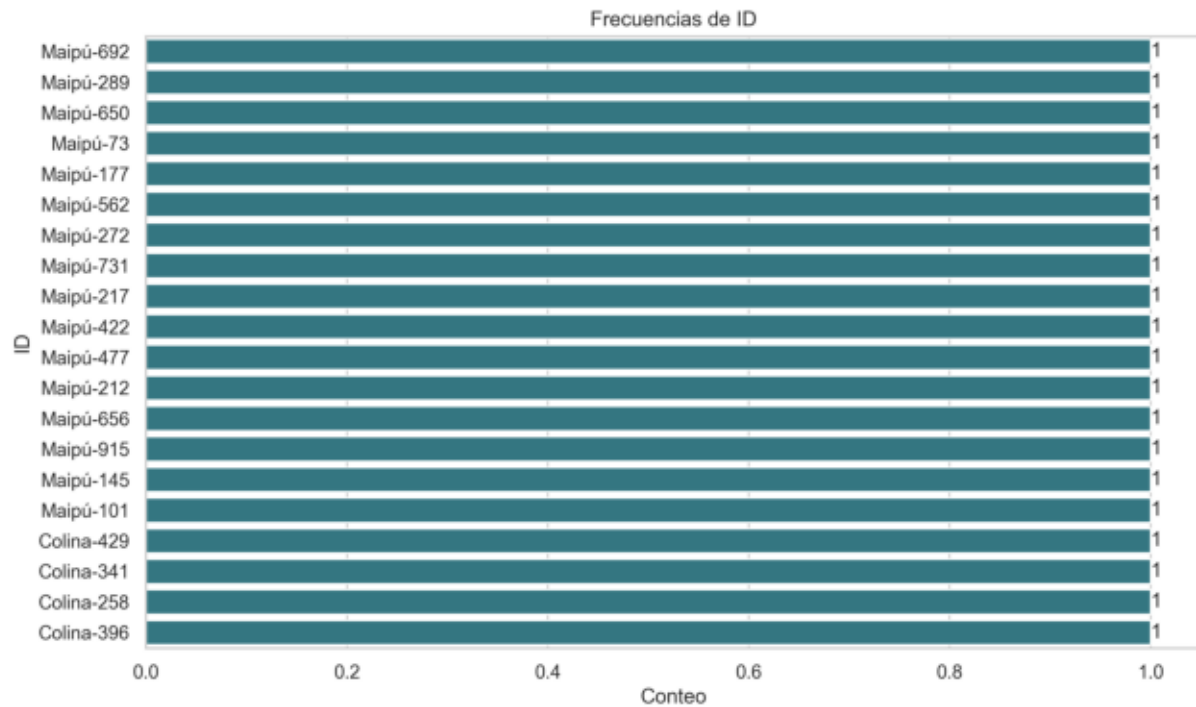
categoricas_barras_Calle.png



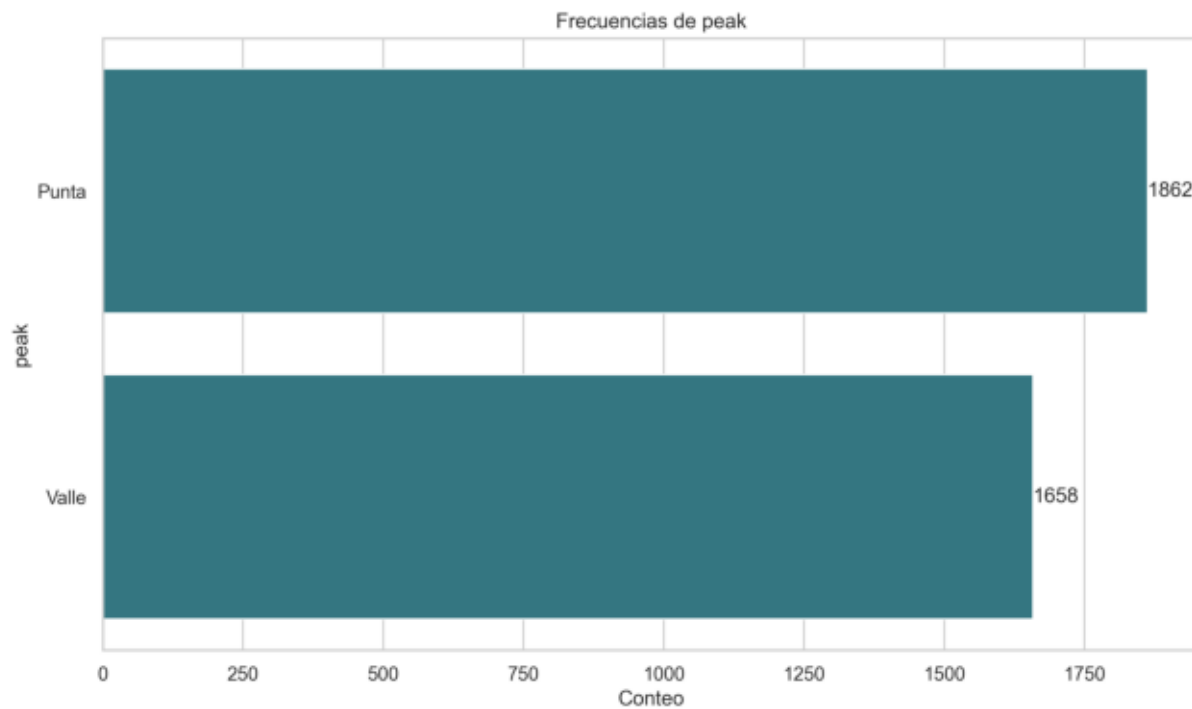
categoricas_barras_Comuna.png



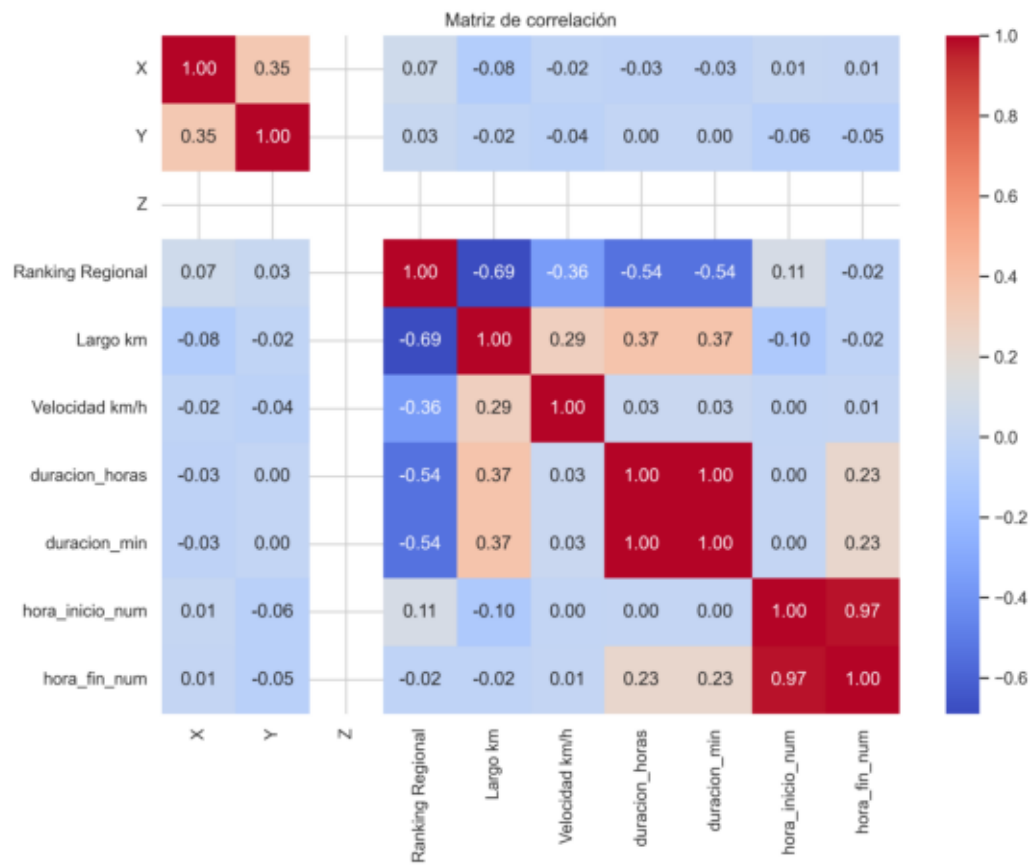
categoricas_barras_ID.png



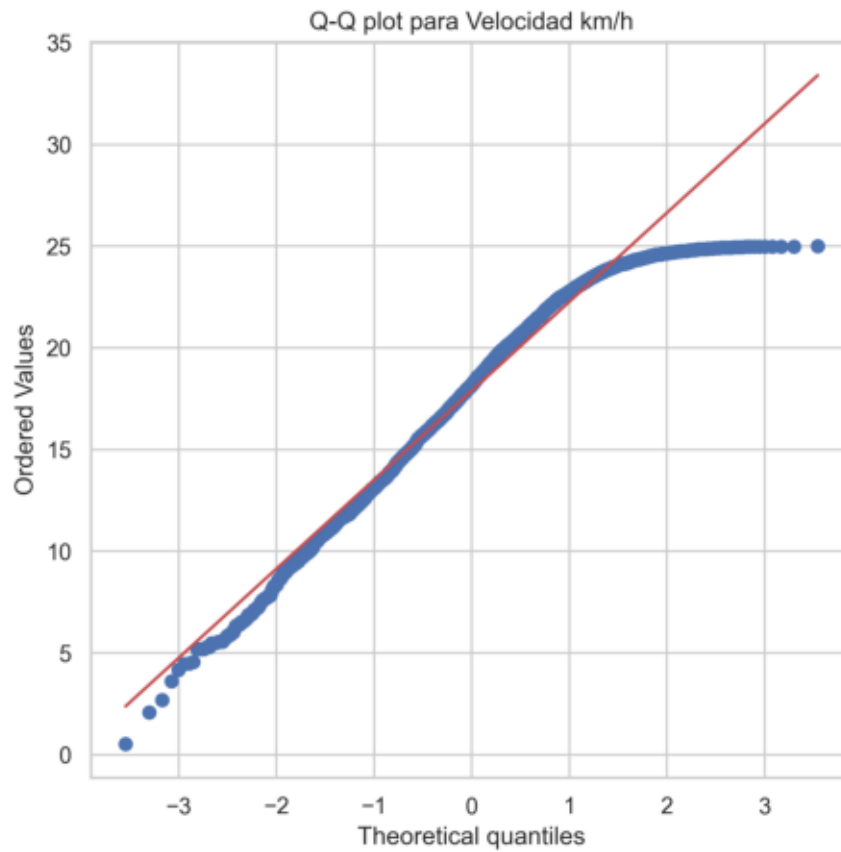
categoricas_barras_peak.png



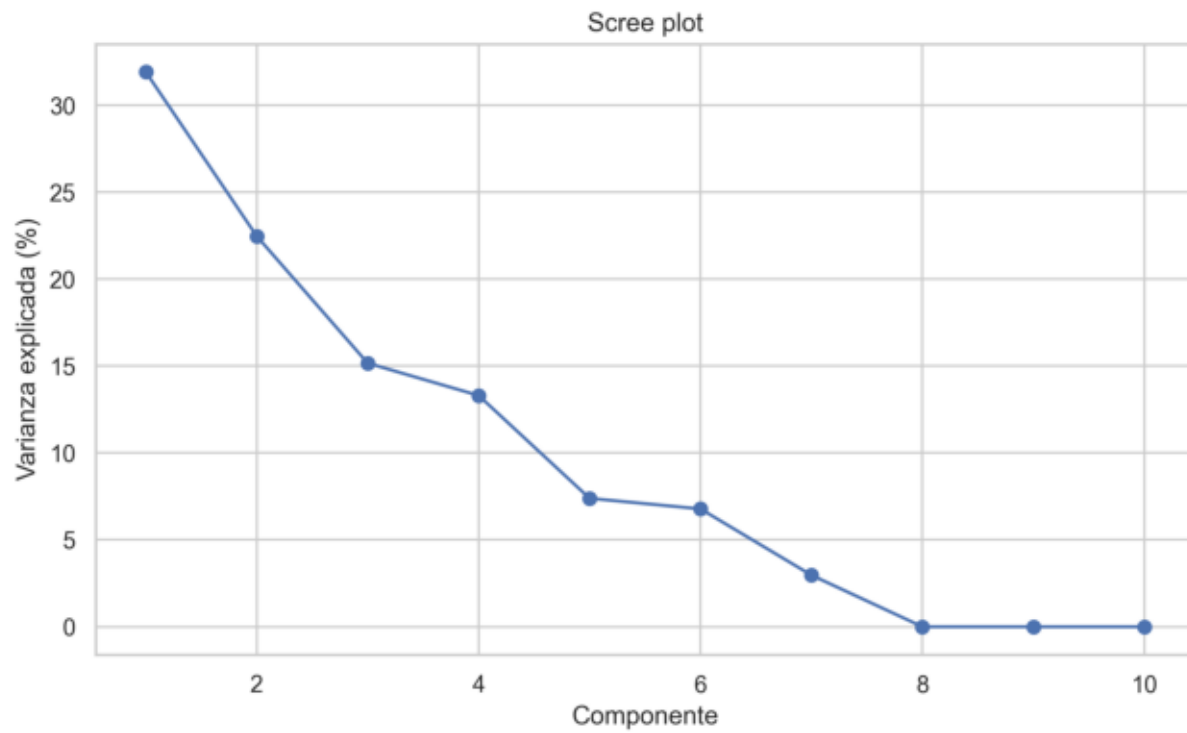
correlacion_heatmap.png



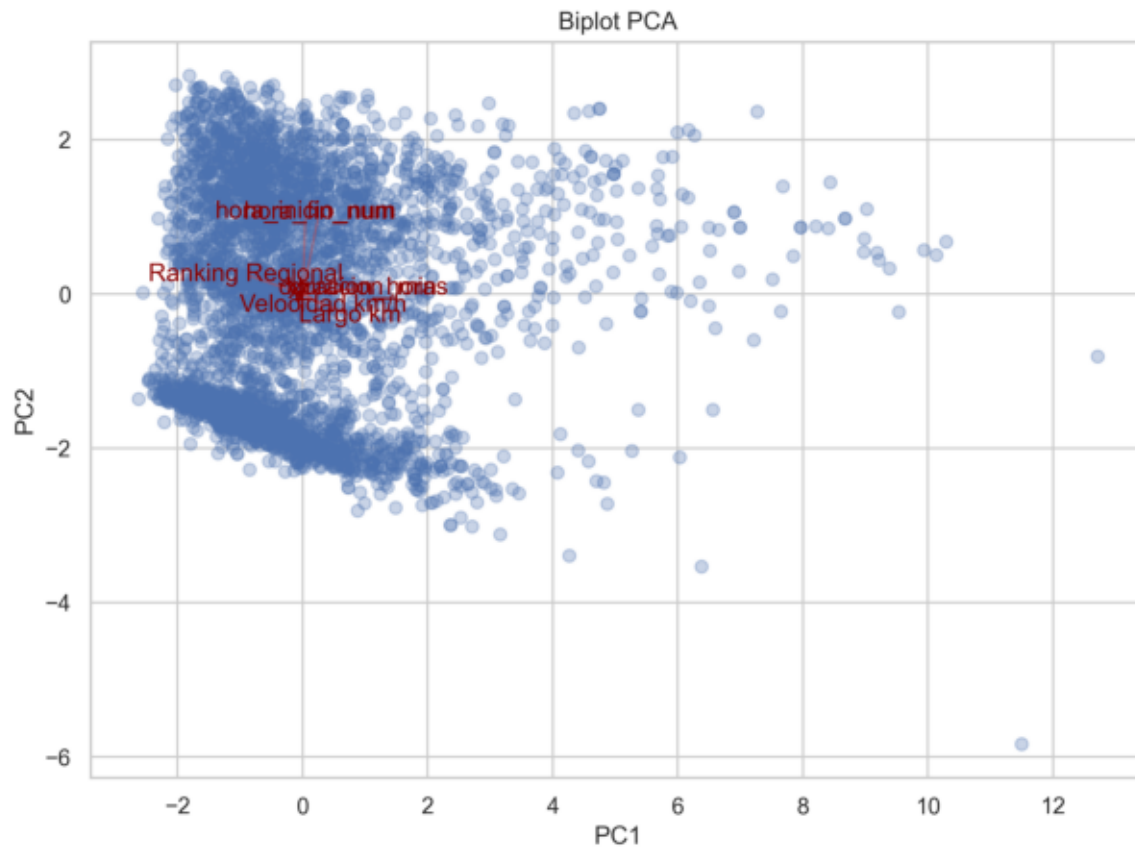
qqplot_Velocidad_kmh.png



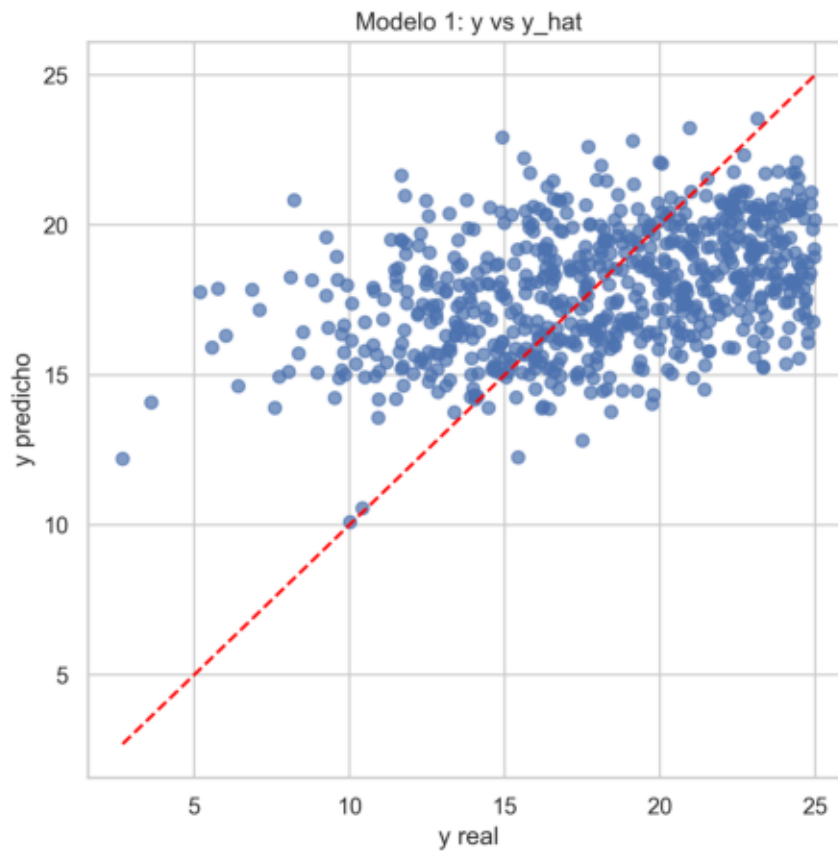
pca_varianza.png



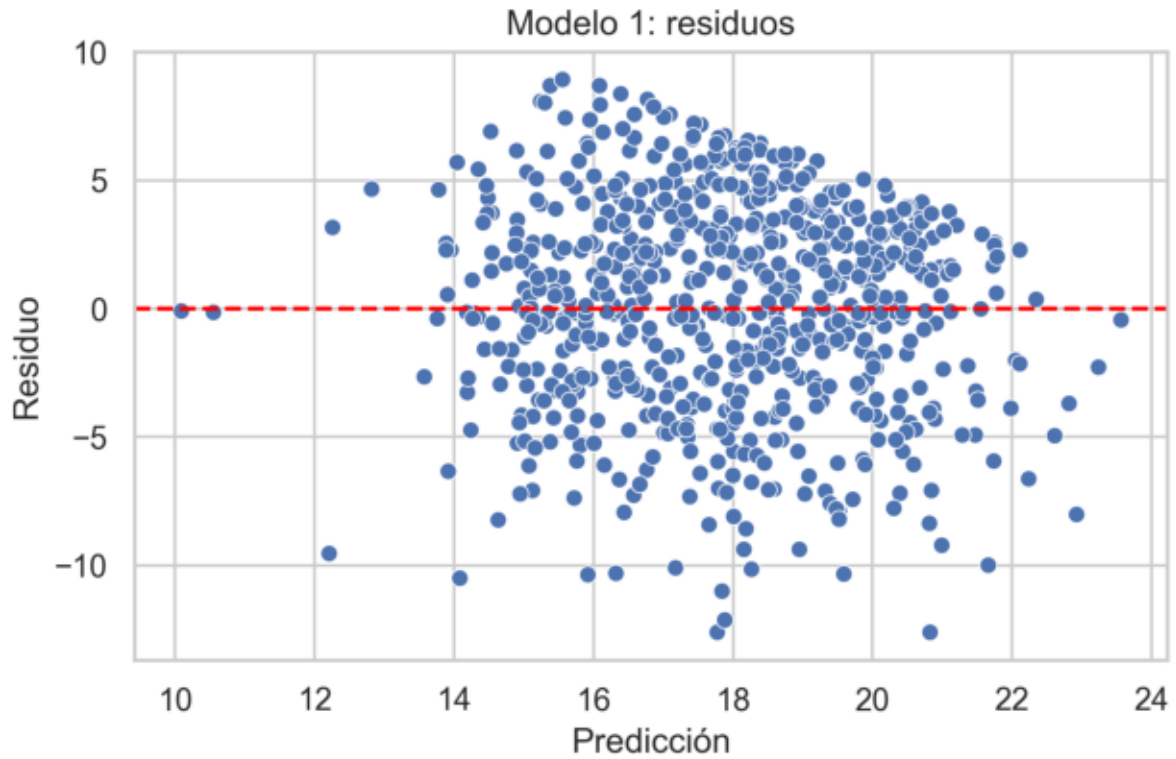
pca_biplot.png



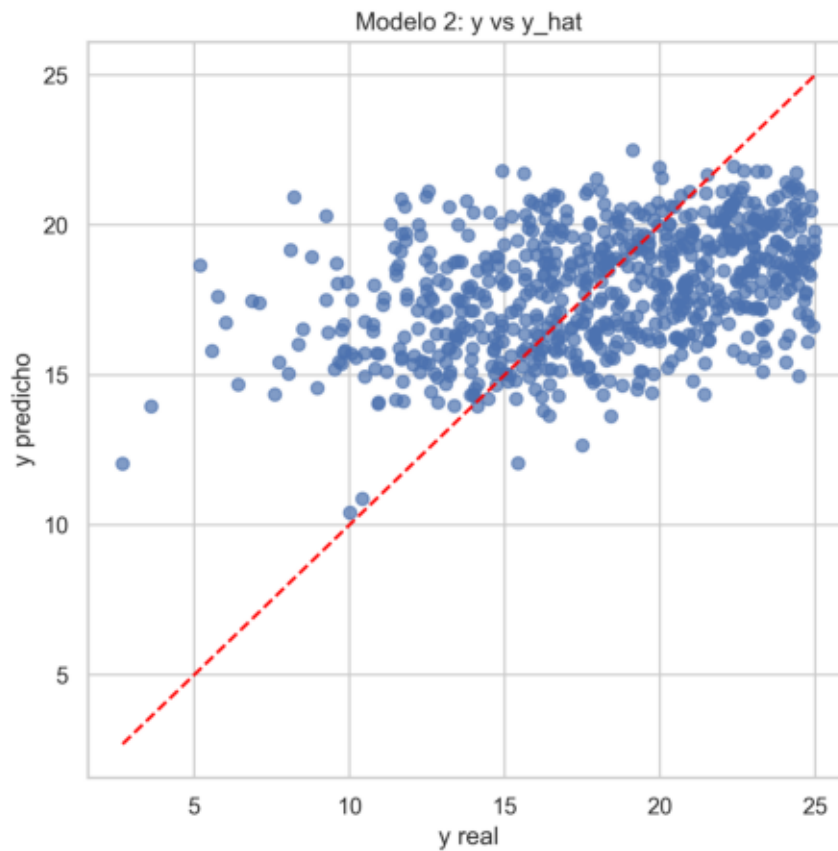
modelo_scatter_y_vs_yhat_m1.png



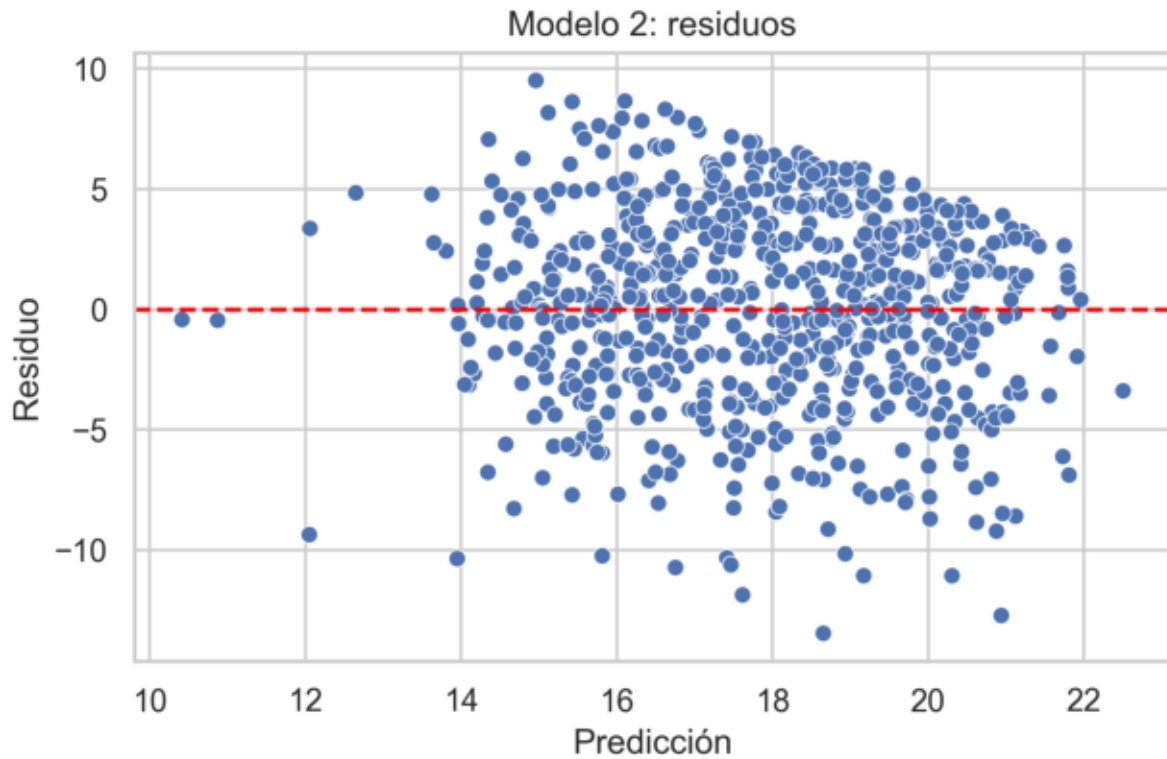
residuos_m1.png



modelo_scatter_y_vs_yhat_m2.png



residuos_m2.png



Conclusiones

Hallazgos:

- Modelo 1 resalta a Comuna_Pudahuel con coeficiente -2.291 (p=nan).
- Modelo 1 resalta a Ranking Regional con coeficiente -2.059 (p=nan).
- Modelo 1 resalta a Comuna_Melipilla con coeficiente 1.673 (p=nan).
- Modelo 1 resalta a Comuna_Recoleta con coeficiente -1.311 (p=nan).
- Modelo 1 resalta a Comuna_Peñalolén con coeficiente -1.275 (p=nan).
- Modelo 2 identifica a Comuna_Pudahuel con coeficiente -2.432 (p=0.000).
- Modelo 2 identifica a Ranking Regional con coeficiente -2.272 (p=0.000).
- Modelo 2 identifica a Comuna_La Pintana con coeficiente 1.439 (p=0.047).
- Modelo 2 identifica a Comuna_Recoleta con coeficiente -1.109 (p=0.016).
- Modelo 2 identifica a duracion_horas con coeficiente -1.078 (p=0.000).
- Correlación destacada entre duracion_horas y duracion_min: 1.00.
- Correlación destacada entre hora_inicio_num y hora_fin_num: 0.97.
- Correlación destacada entre Ranking Regional y Largo km: -0.69.
- Correlación destacada entre Ranking Regional y duracion_min: -0.54.

Referencias

Dataset: Observatorio de Transporte, congestión en Santiago (14/03/2025).

Herramientas: pandas, numpy, matplotlib, seaborn, scikit-learn, statsmodels.