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 20 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 7 componentes (hasta PC7) 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: 4.065, R2: 0.158, MAE: 3.324, Bias: -0.158.

Modelo 2

Modelo lineal con selección secuencial de variables para mejorar interpretabilidad.

Desempeño: RMSE: 4.067, R2: 0.157, MAE: 3.312, Bias: -0.145.

Coeficientes Destacados

Modelo 1 (fuente: ols):

- Comuna_Pudahuel: coef=-2.956 |coef|=2.956

- Comuna_Recoleta: coef=-2.086 |coef|=2.086

- Comuna_Peñalolén: coef=-1.885 |coef|=1.885

- Comuna_Lampa: coef=-1.789 |coef|=1.789

- Comuna_Quinta Normal: coef=-1.774 |coef|=1.774

- Comuna_Independencia: coef=-1.765 |coef|=1.765

- Comuna_La Reina: coef=-1.722 |coef|=1.722

- Comuna_Huechuraba: coef=-1.673 |coef|=1.673

- Ranking Regional: coef=-1.466 |coef|=1.466

- hora_central: coef=1.372 |coef|=1.372

Modelo 2 (fuente: ols):

- Comuna_Pudahuel: coef=-2.213 |coef|=2.213

- Comuna_La Pintana: coef=1.585 |coef|=1.585

- Ranking Regional: coef=-1.480 |coef|=1.480

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
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
Largo_km_log1p	3520.0	0.32	0.16	0.11	0.21	0.28	0.38	2.23	2.29	11.22
Velocidad km/h	3520.0	17.89	4.43	0.53	14.79	18.2	21.49	25.0	-0.44	-0.4
duracion_min	3520.0	49.79	59.97	15.0	19.8	30.0	55.2	555.0	3.86	19.3
hora_central	3520.0	13.54	4.32	6.38	8.38	13.88	17.42	21.46	-0.06	-1.42
hora_sin	3520.0	-0.19	0.77	-1.0	-0.94	-0.47	0.81	1.0	0.4	-1.58
hora_cos	3520.0	-0.42	0.44	-1.0	-0.77	-0.5	-0.15	0.79	0.75	-0.13

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
Calle	39	1.11
Comuna	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_min	0	0.0
peak	0	0.0
hora_central	0	0.0
hora_sin	0	0.0
hora_cos	0	0.0
TARGET_BIN	0	0.0
Largo_km_log1p	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
hora_central	hora_sin	-0.9613774196303412
Largo km	Largo_km_log1p	0.94905117828908
Ranking Regional	Largo_km_log1p	-0.8274164021335929
Ranking Regional	Largo km	-0.6897624194106322
hora_central	hora_cos	0.6290916828576741
Ranking Regional	duracion_min	-0.5406030400573674

tabla pca varianza

componente	var_exp	var_exp_acum
PC1	31.17	31.17
PC2	24.02	55.19
PC3	13.38	68.57
PC4	9.68	78.24
PC5	6.78	85.02
PC6	6.54	91.56
PC7	5.52	97.09
PC8	2.56	99.65
PC9	0.24	99.89
PC10	0.11	100.0

tabla pca cargas pc1

variable	loading	abs_loading_rank
Largo_km_log1p	0.5440812052226512	1
Ranking Regional	-0.5102337255313737	2
Largo km	0.5078492608492895	3
duracion_min	0.3228577666661482	4
Velocidad km/h	0.2572036456273345	5
hora_central	-0.0659374471982627	6
X	-0.0656463802111209	7
hora_sin	0.0581237371288298	8
hora_cos	-0.0508418104311667	9
Y	-0.0267396747257938	10

tabla pca cargas pc2

variable	loading	abs_loading_rank
hora_central	0.6295464072582981	1
hora_sin	-0.5912174499051844	2
hora_cos	0.4716295728640191	3
duracion_min	0.1549820324615411	4
Y	-0.0673084528907272	5
Ranking Regional	-0.0428453082624744	6
Velocidad km/h	0.0220044211369315	7
Largo_km_log1p	0.0191984983383136	8
Largo km	0.0183512734562165	9
X	-0.0111206678684606	10

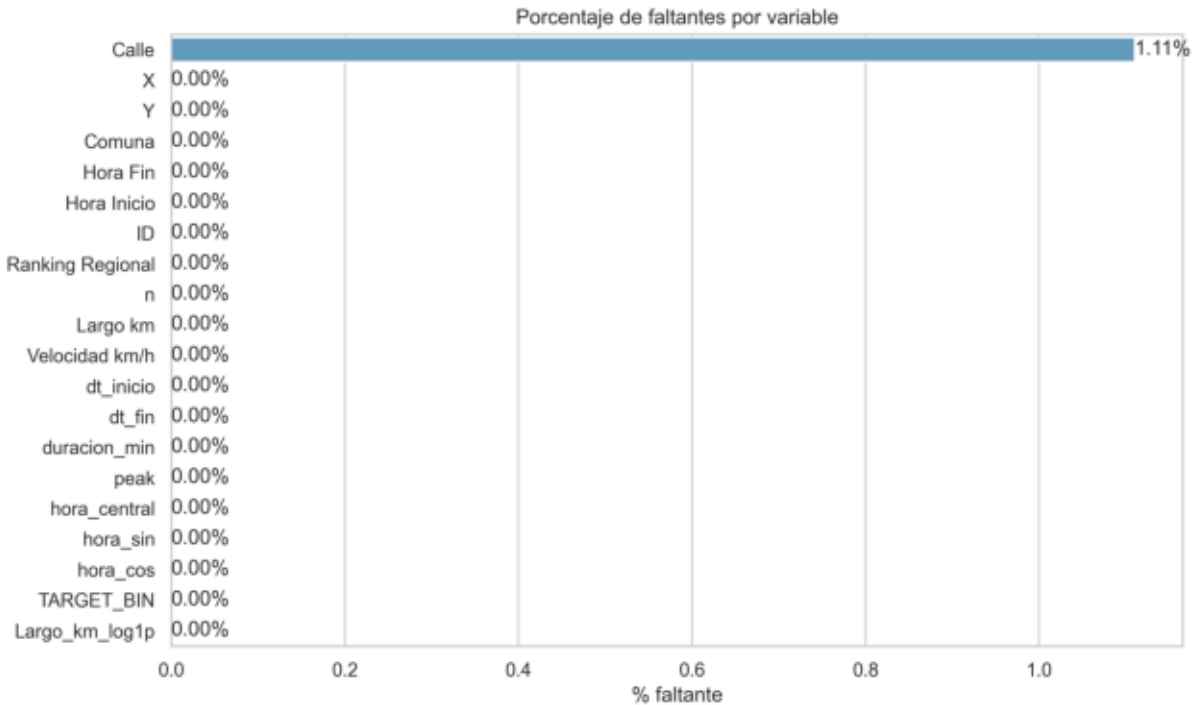
metricas modelos

modelo	RMSE	R2	MAE	Bias
baseline	4.432930176090961	-0.0008677215865517	3.687786232869936	-0.1305247138805066
modelo1	4.064751486858197	0.158482989867234	3.3235226735855408	-0.1581048569804235
modelo2	4.067306718020874	0.1574246489776916	3.3120890512678165	-0.1453209419097414

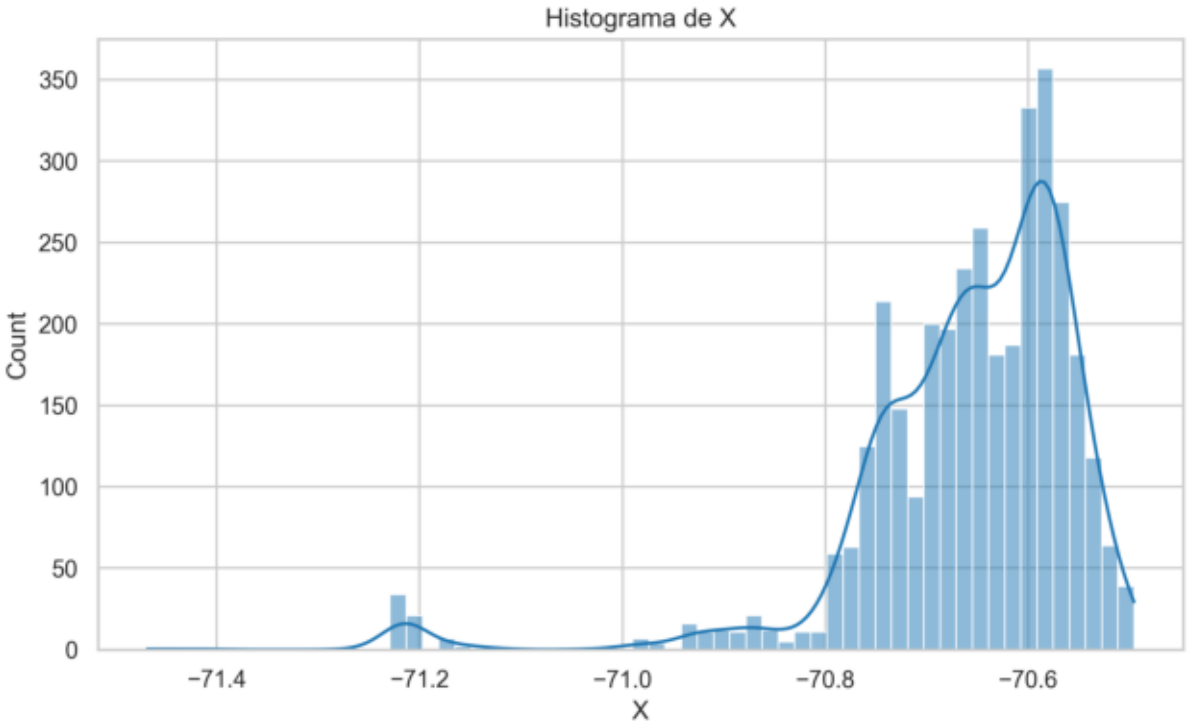
colinearidad vif

variable	VIF
hora_central	50.36088531491082
hora_sin	37.488479675221775
hora_cos	4.716528969395293
Ranking Regional	3.73060204115168
Largo_km_log1p	3.2358487723803524
duracion_min	1.4861847213154966
Y	1.134021229794714
X	1.128538029969349

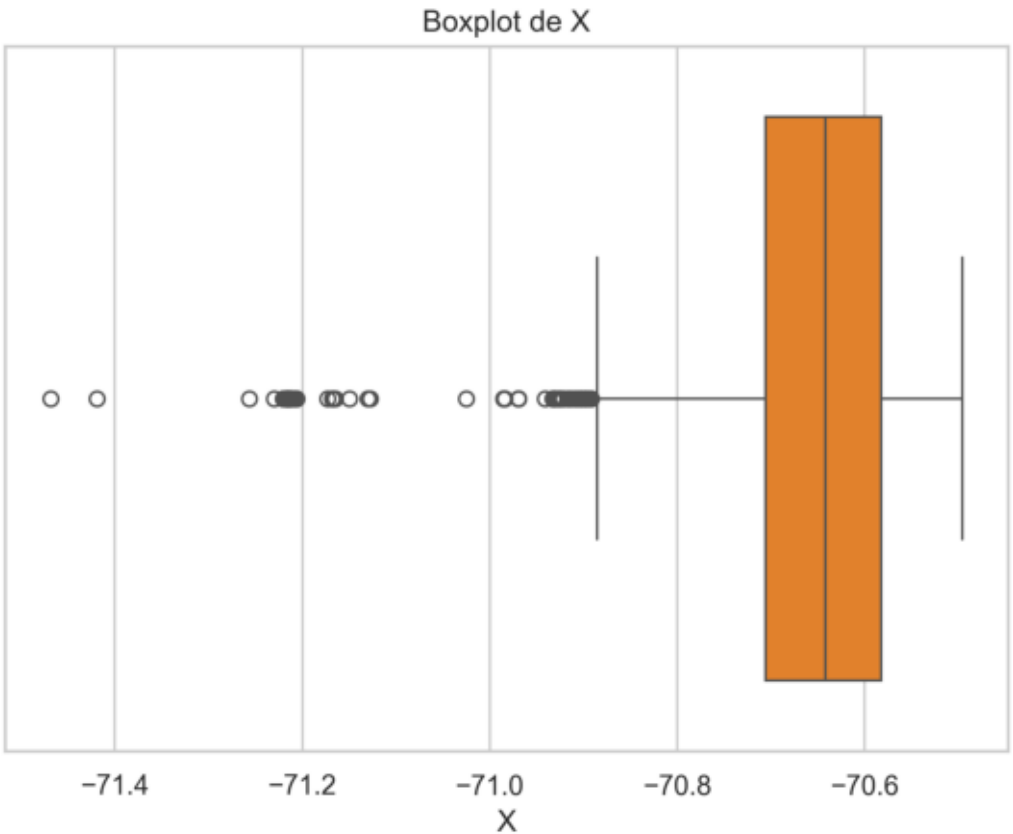
faltantes_bar.png



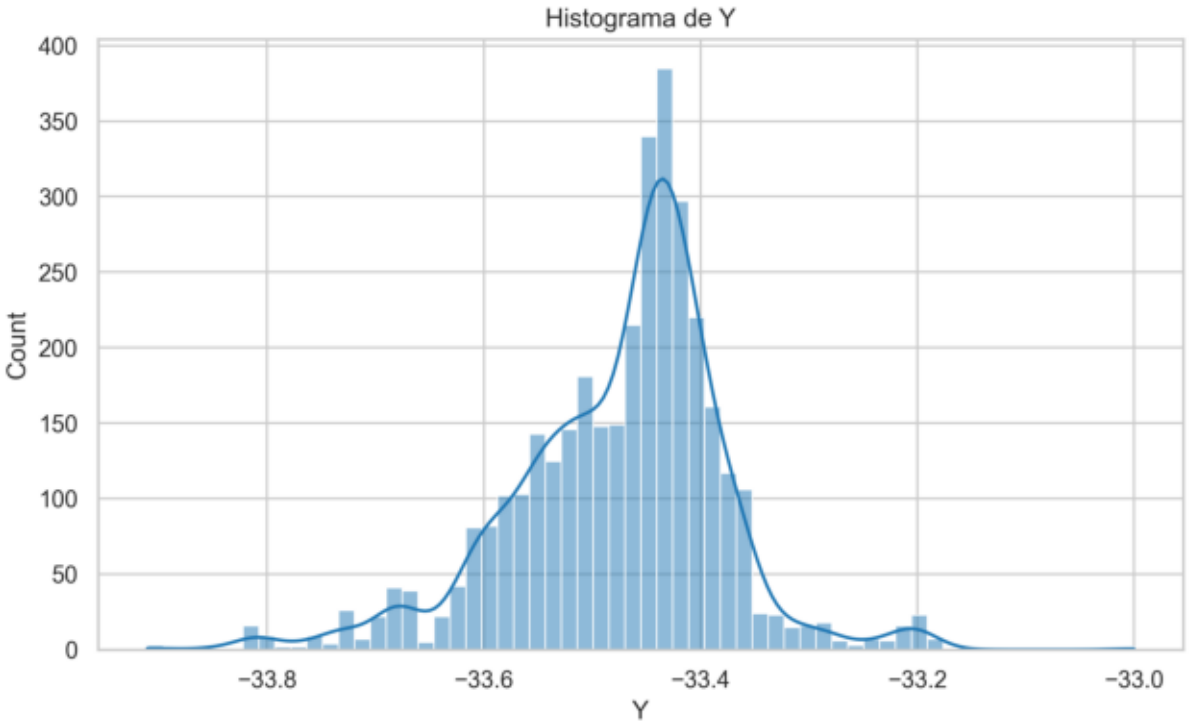
hist_X.png


Histograma de X
Count
350
300
250
200
150
100
50
0
−71.4
−71.2
−71.0
−70.8
−70.6
X

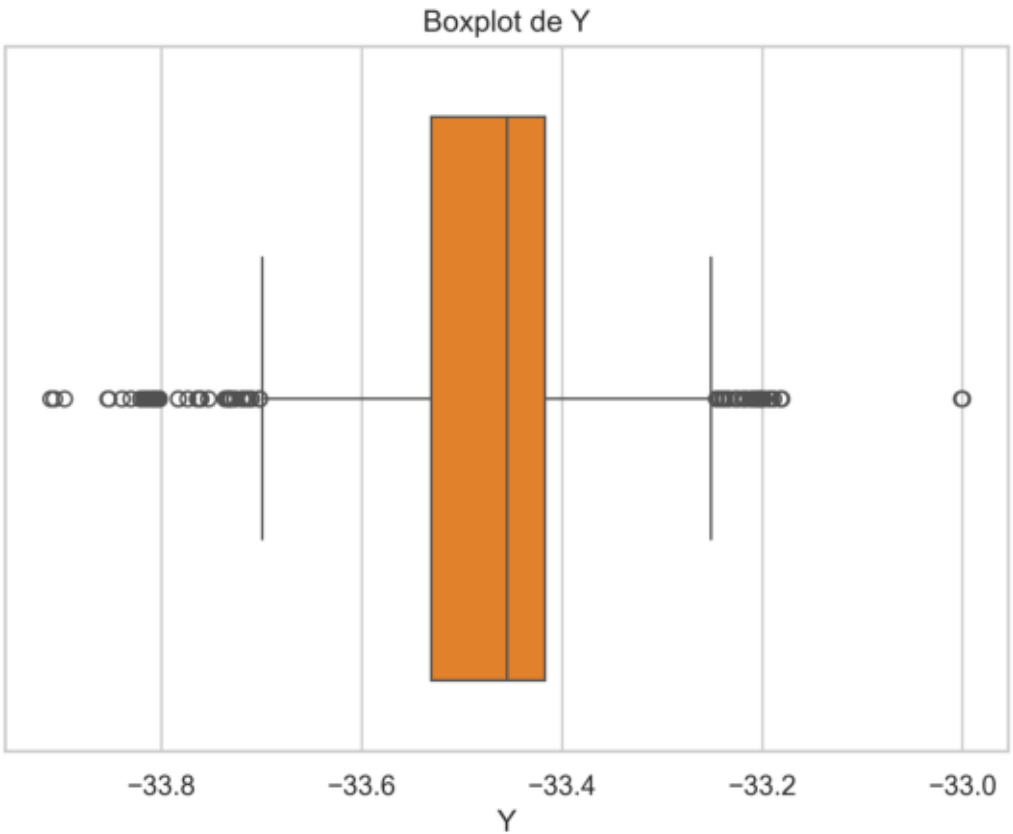
boxplot_X.png



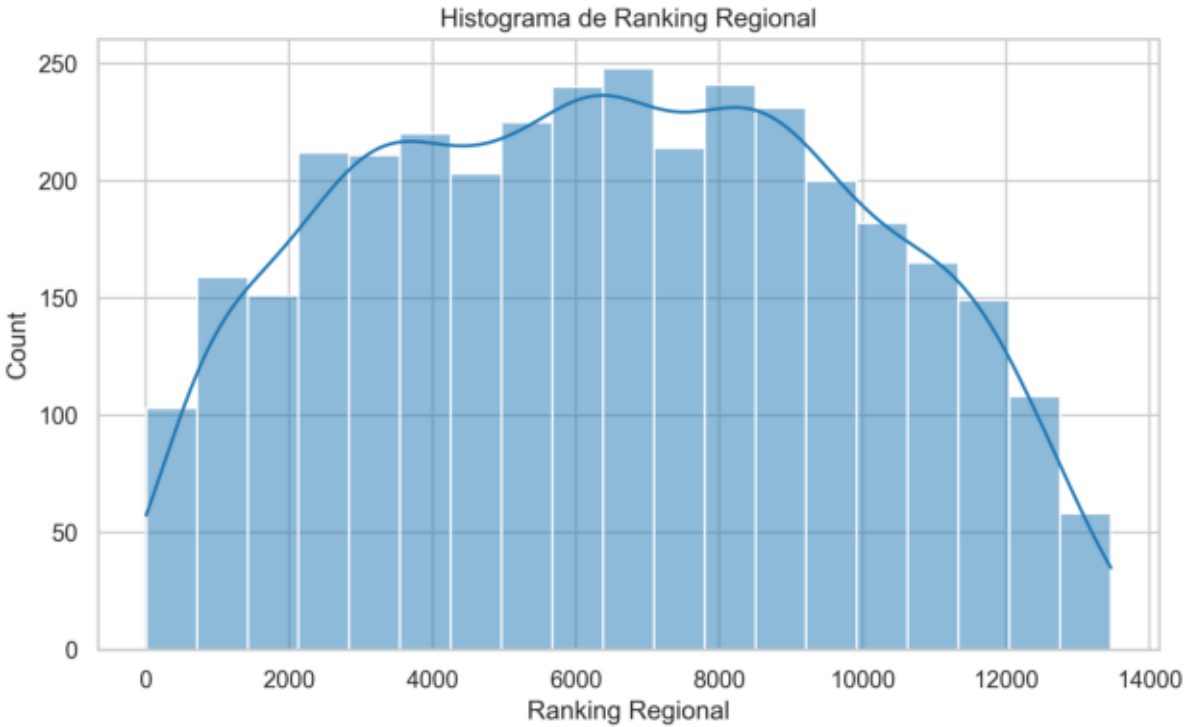
hist_Y.png


Histograma de Y
Count
400
350
300
250
200
150
100
50
0
−33.8
−33.6
−33.4
−33.2
−33.0
Y

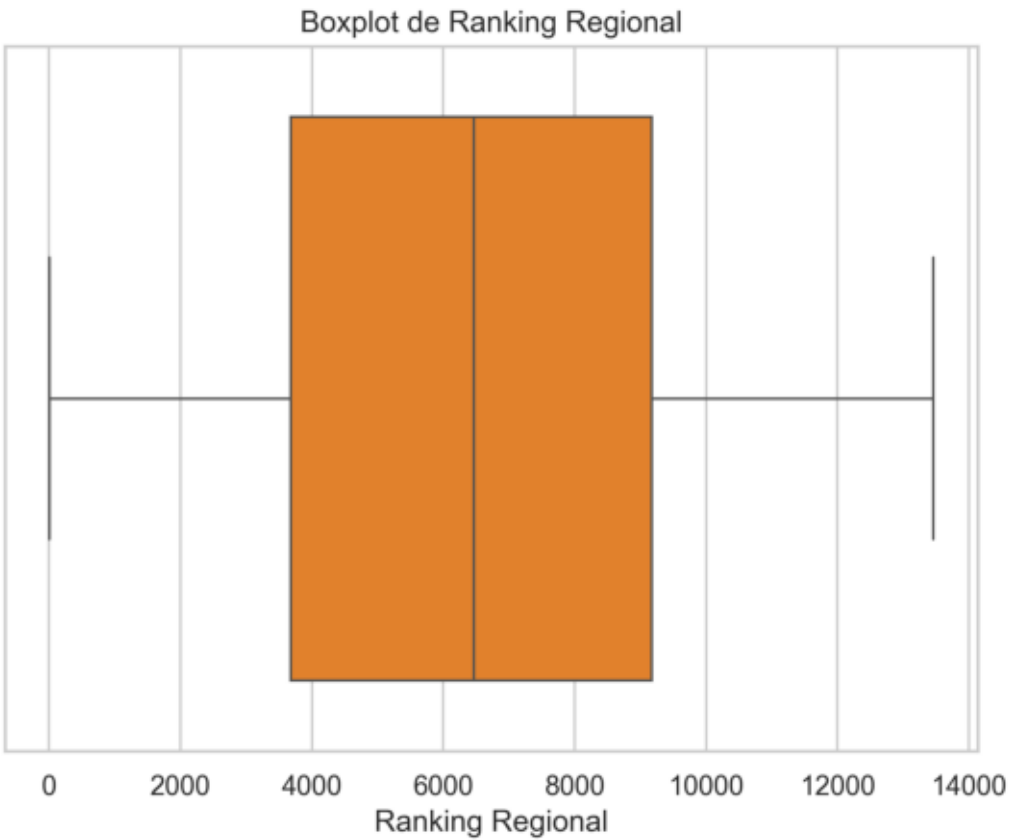
boxplot_Y.png



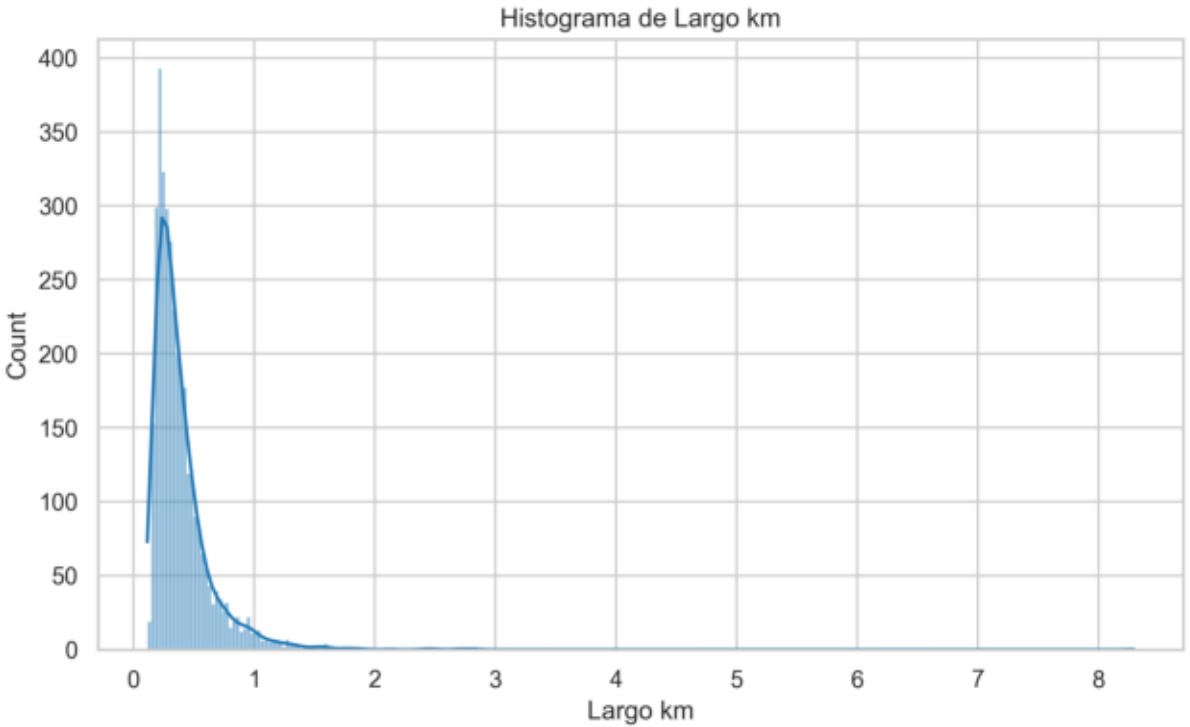
hist_Ranking_Regional.png



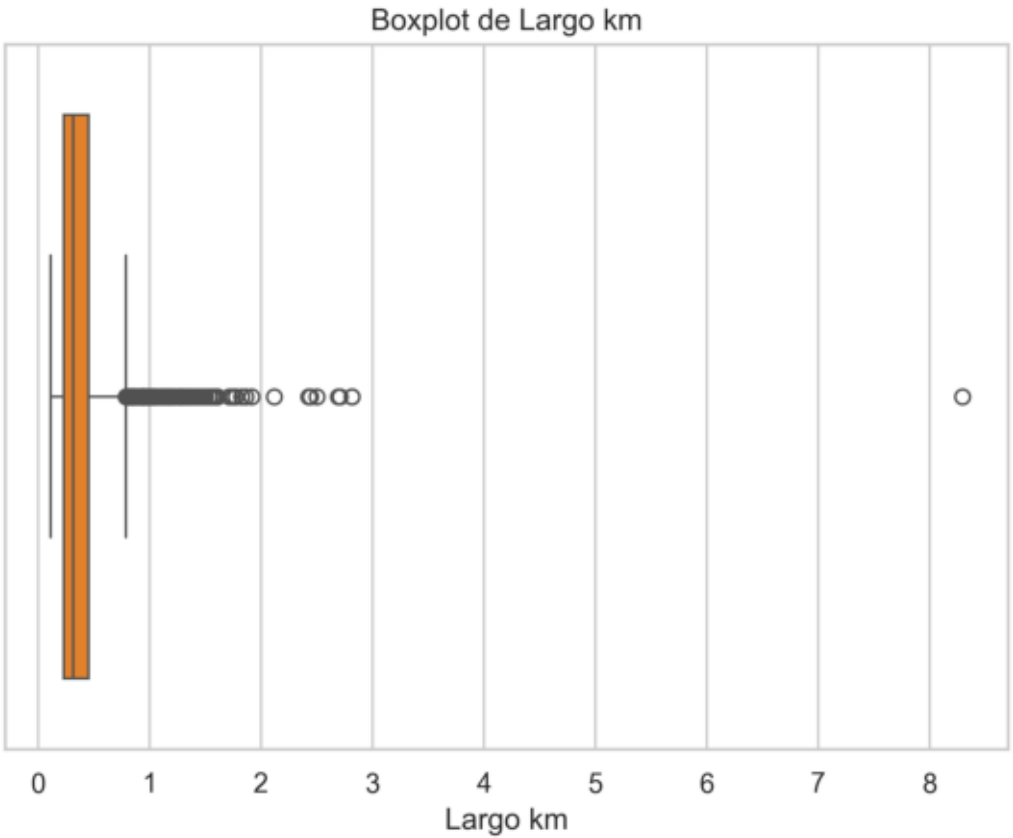
boxplot_Ranking_Regional.png



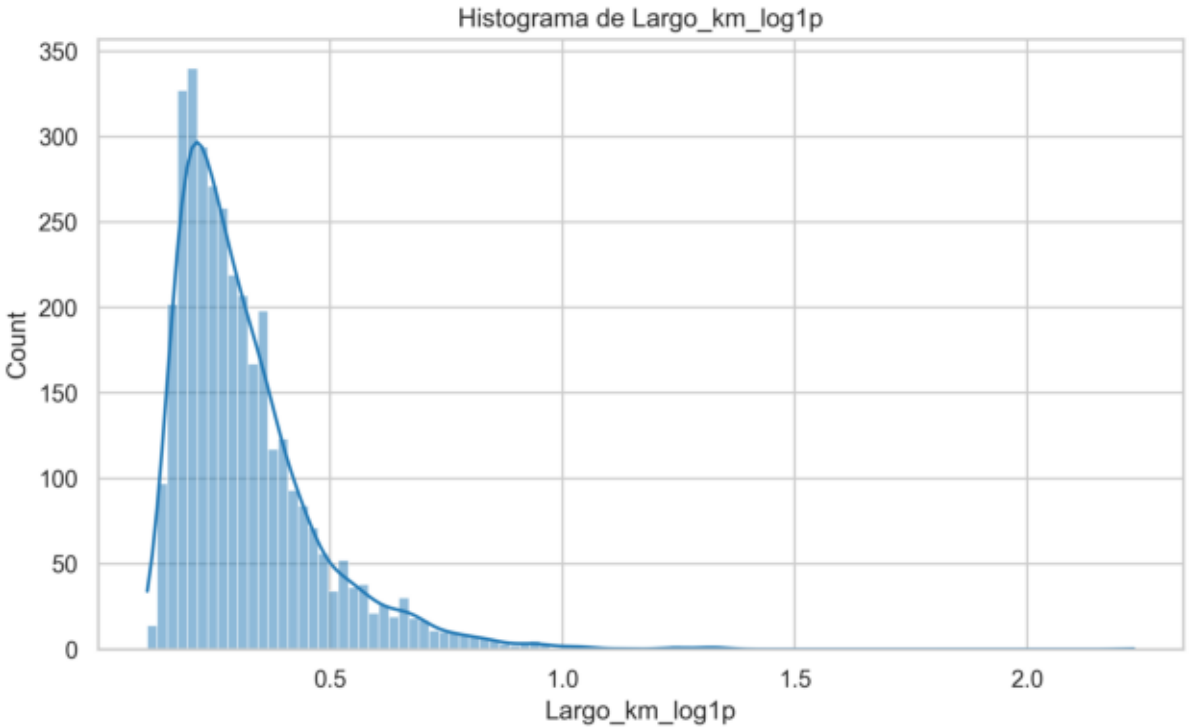
hist_Largo_km.png



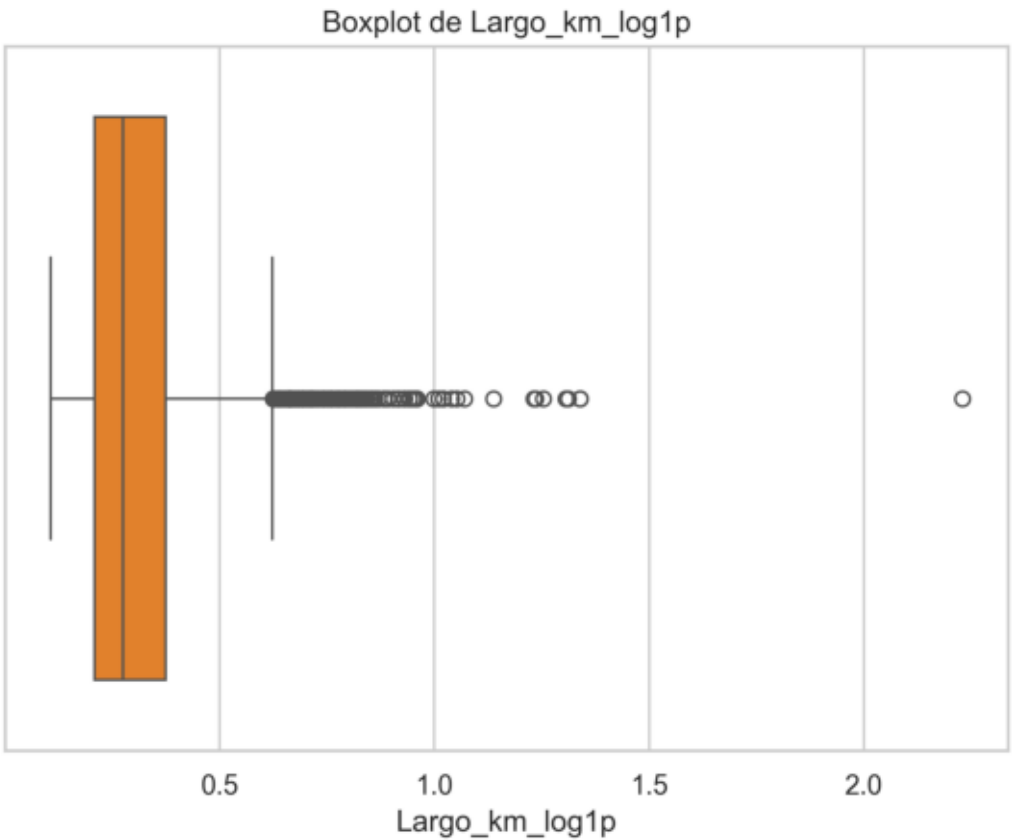
boxplot_Largo_km.png



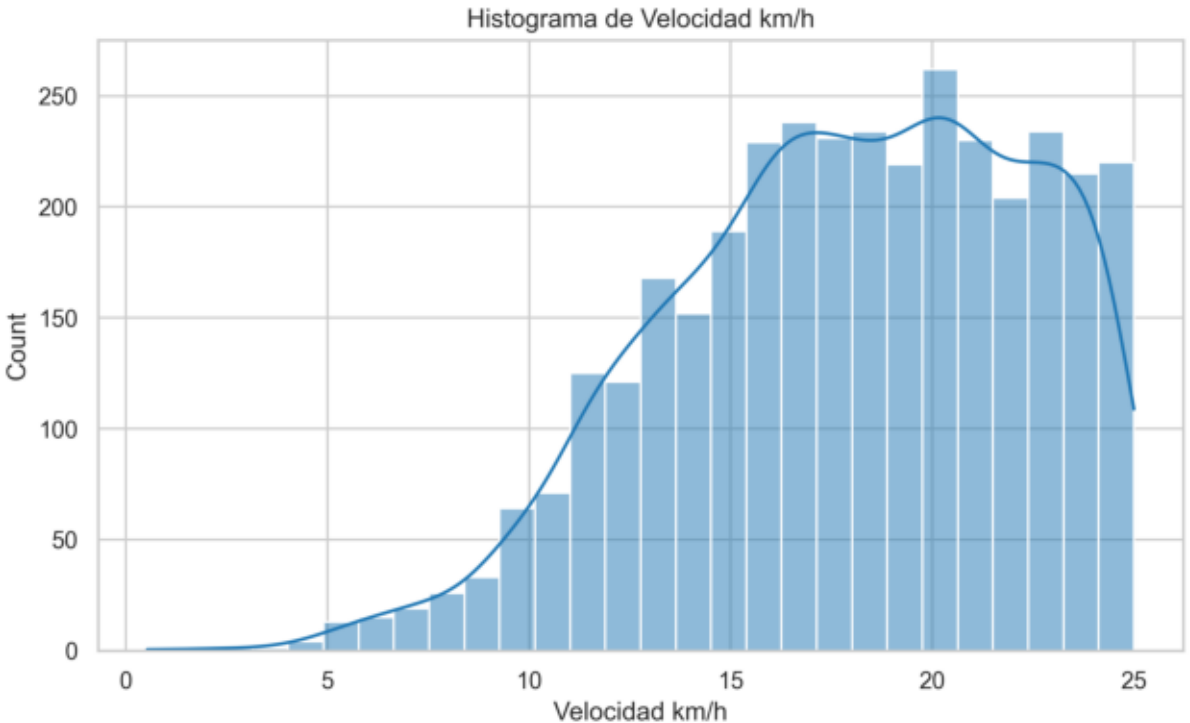
hist_Largo_km_log1p.png



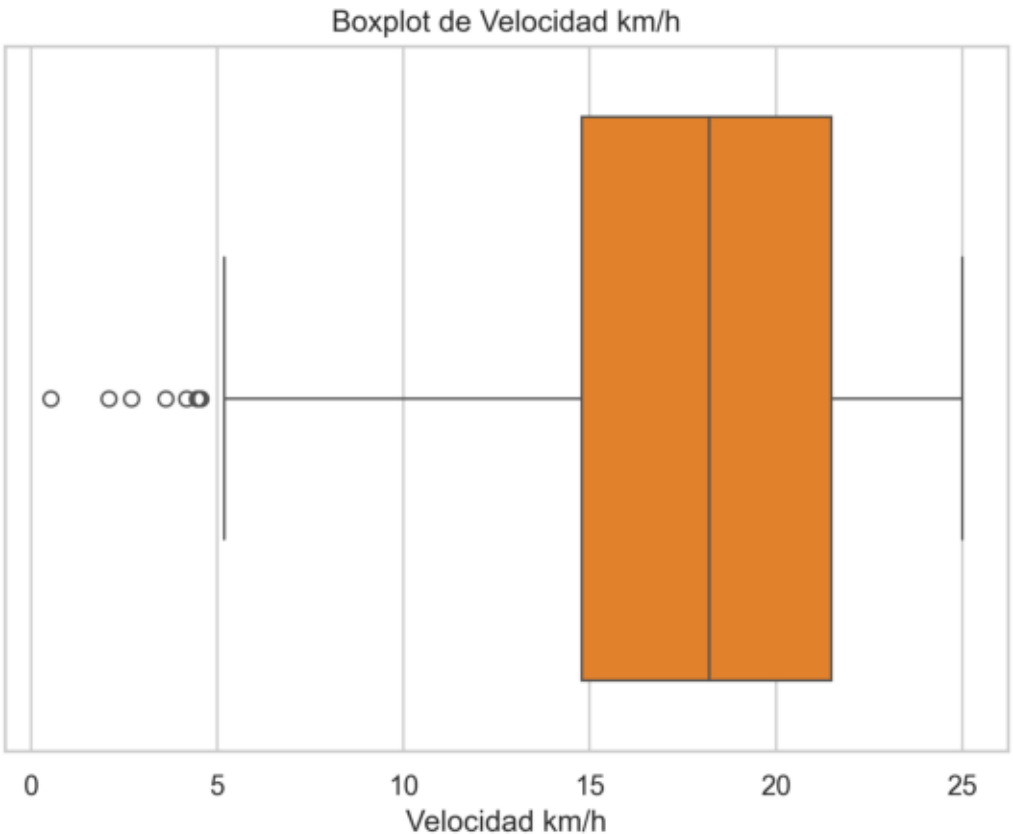
boxplot_Largo_km_log1p.png



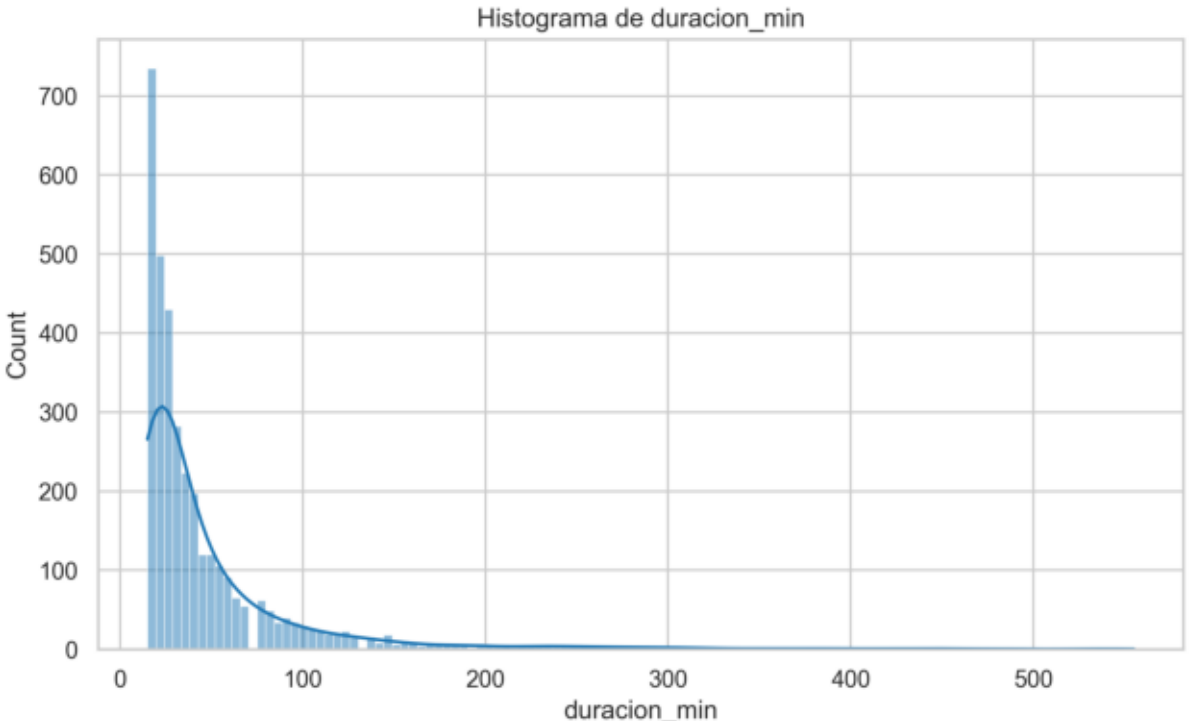
hist_Velocidad_kmh.png



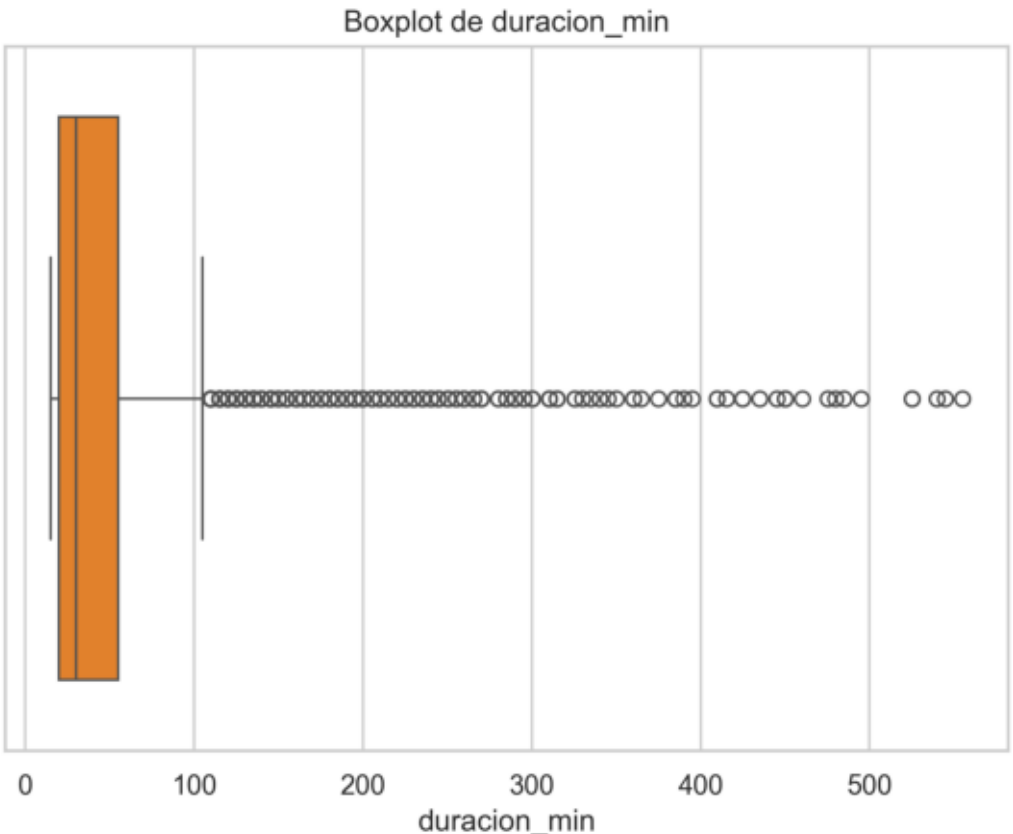
boxplot_Velocidad_kmh.png



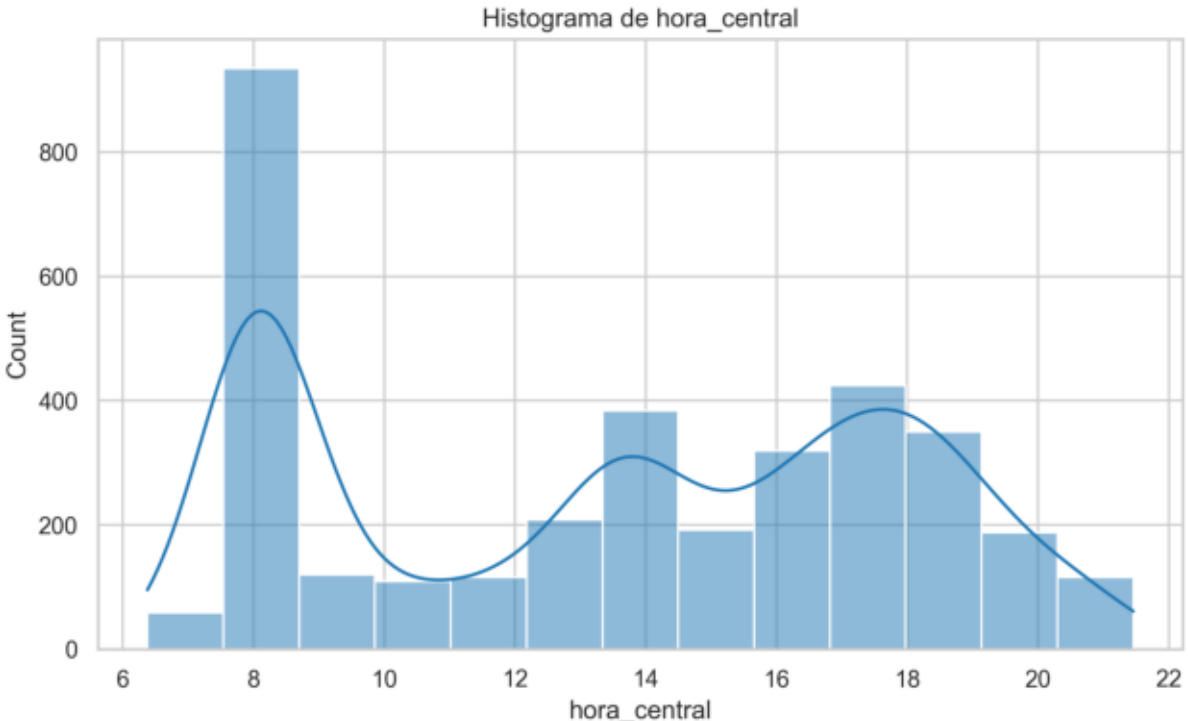
hist_duracion_min.png



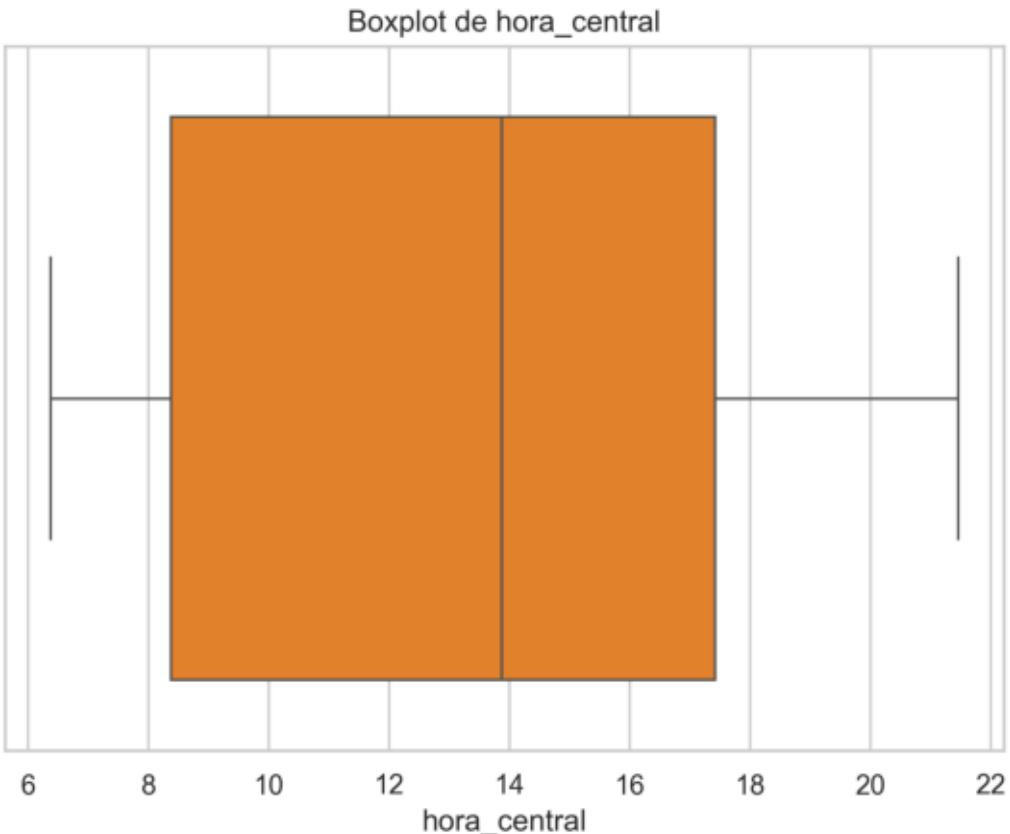
boxplot_duracion_min.png



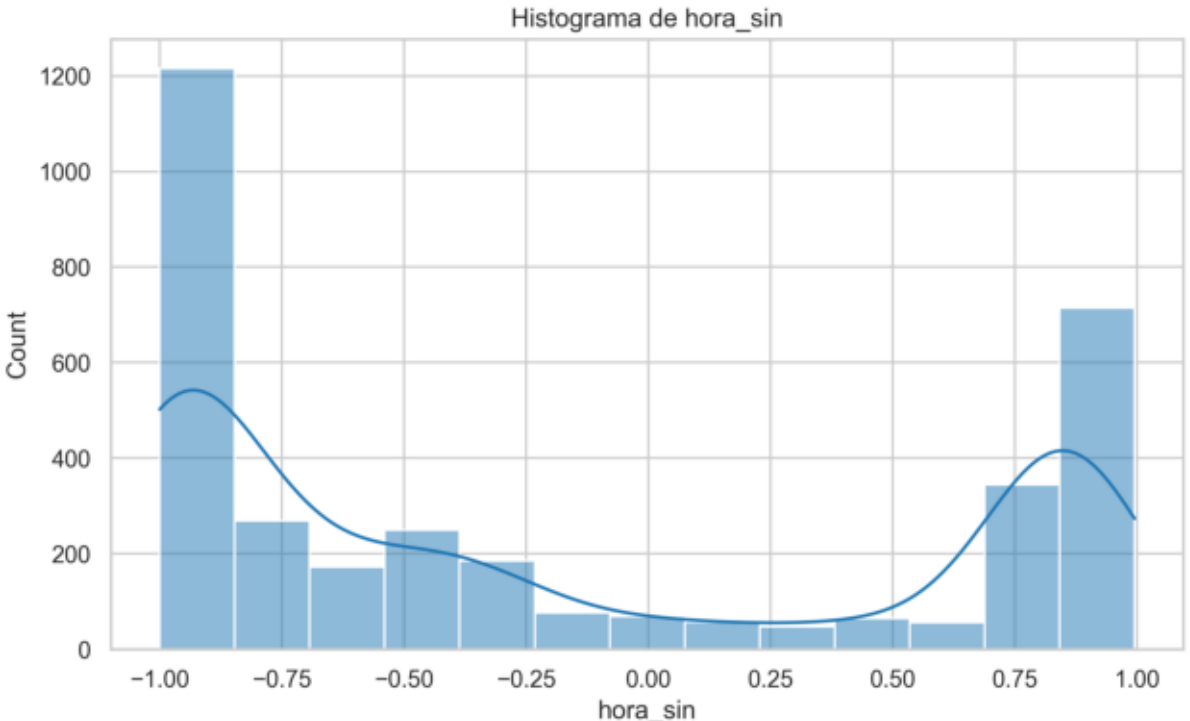
hist_hora_central.png



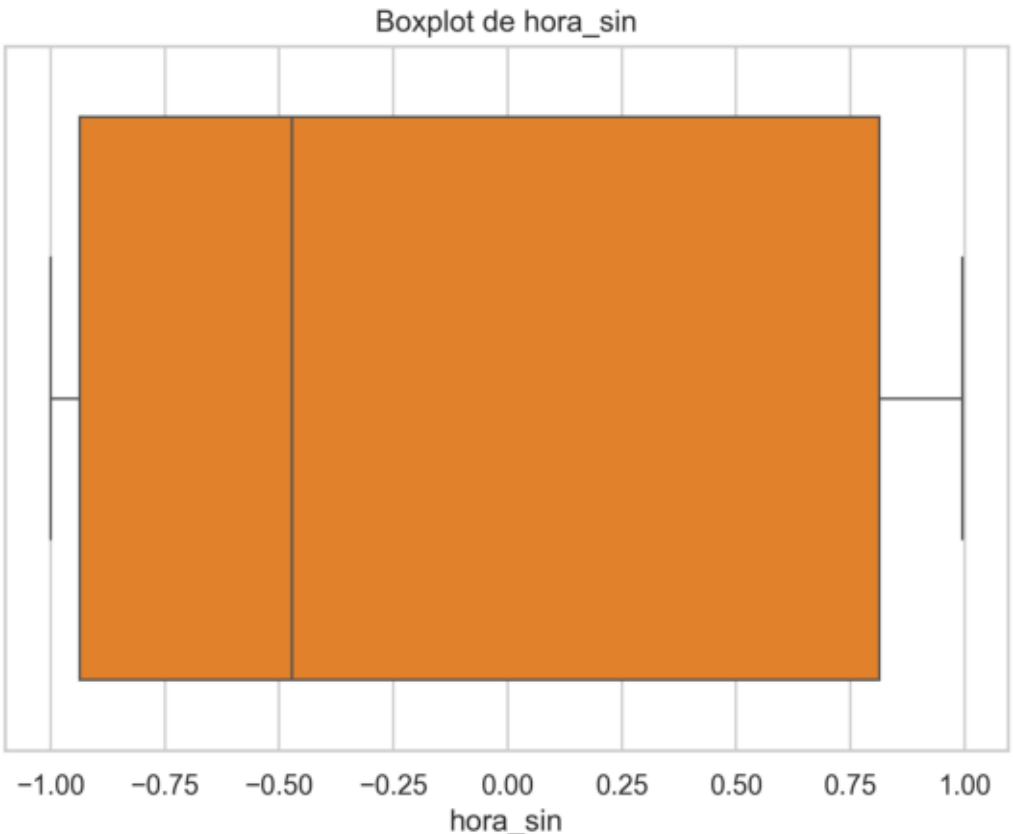
boxplot_hora_central.png



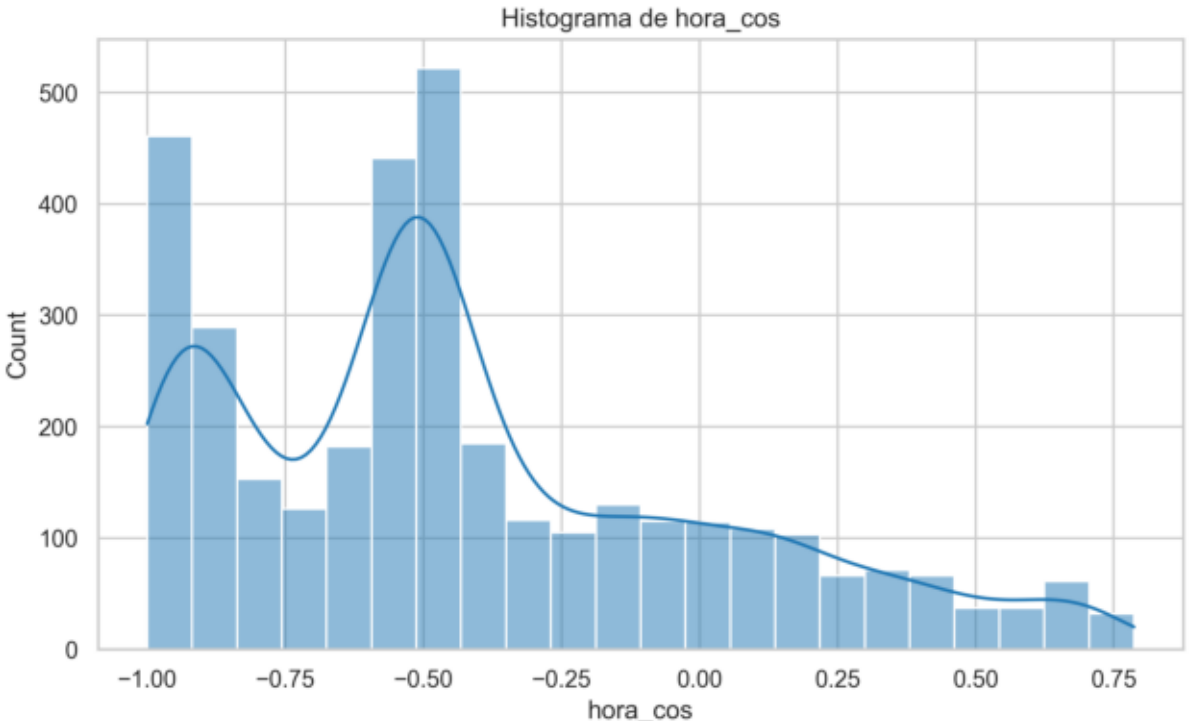
hist_hora_sin.png



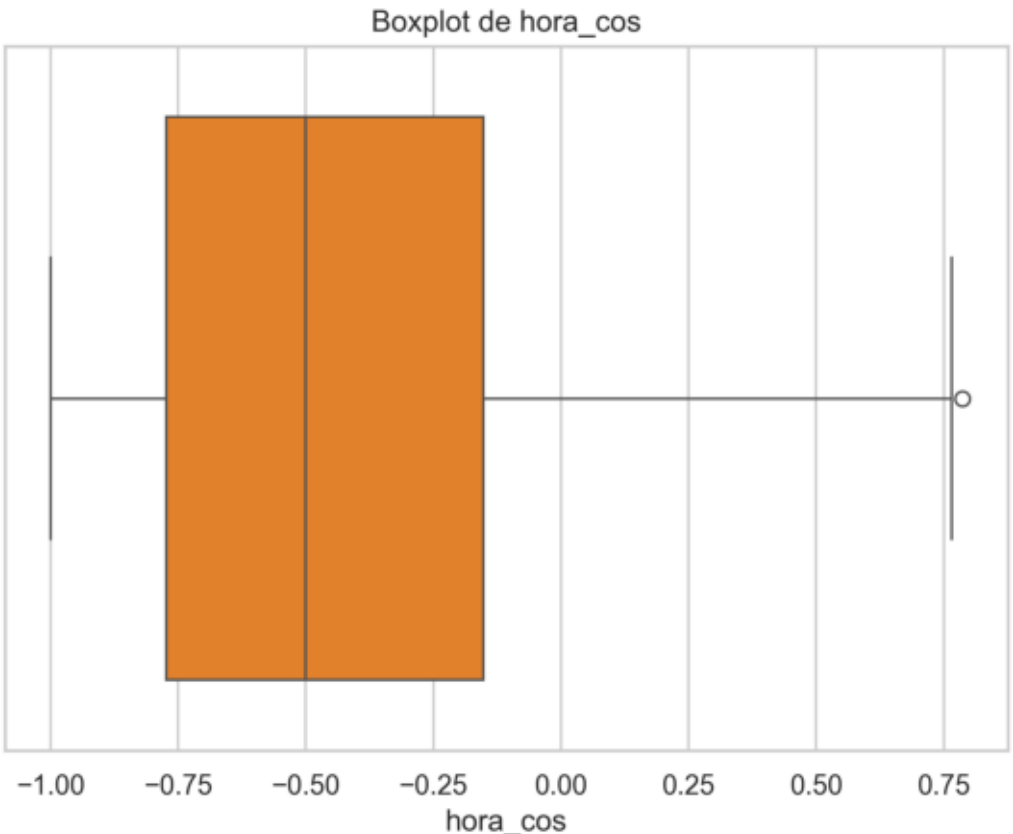
boxplot_hora_sin.png



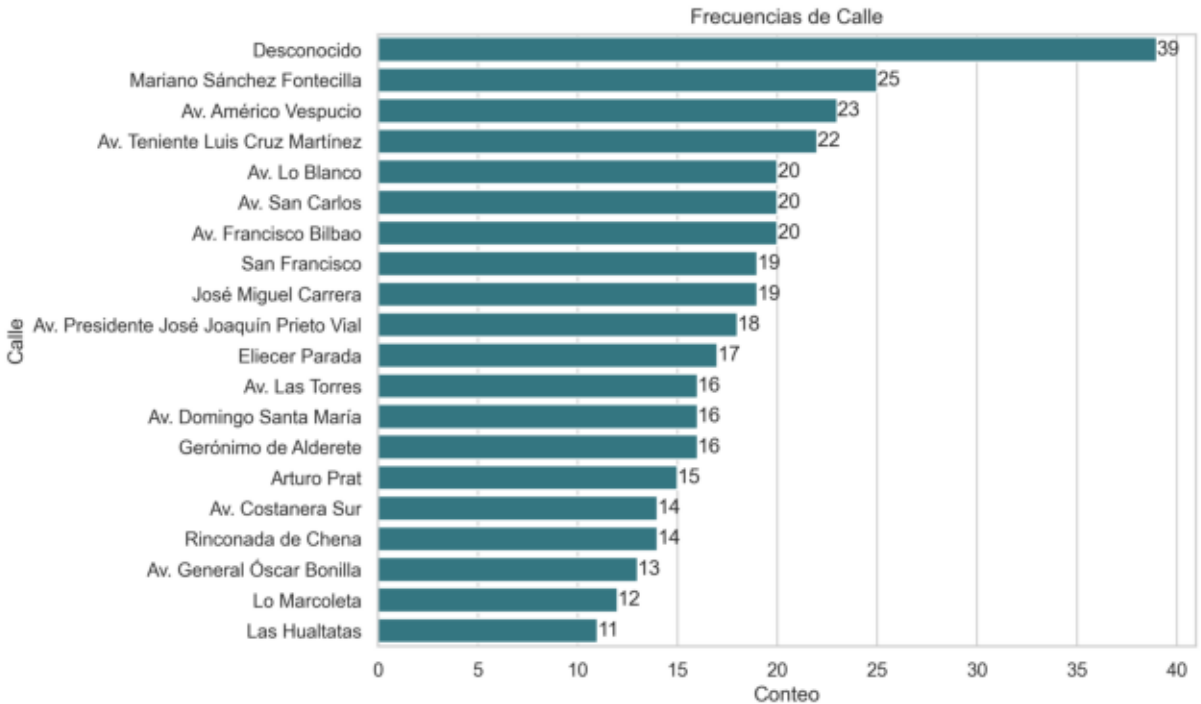
hist_hora_cos.png



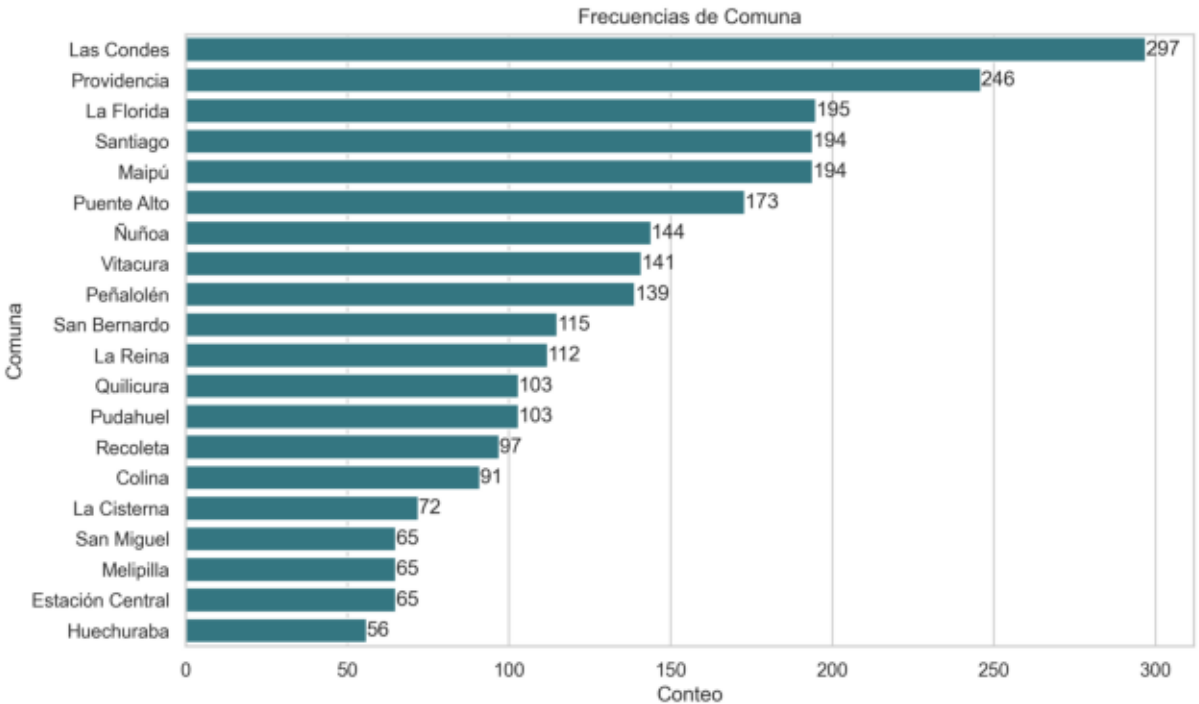
boxplot_hora_cos.png



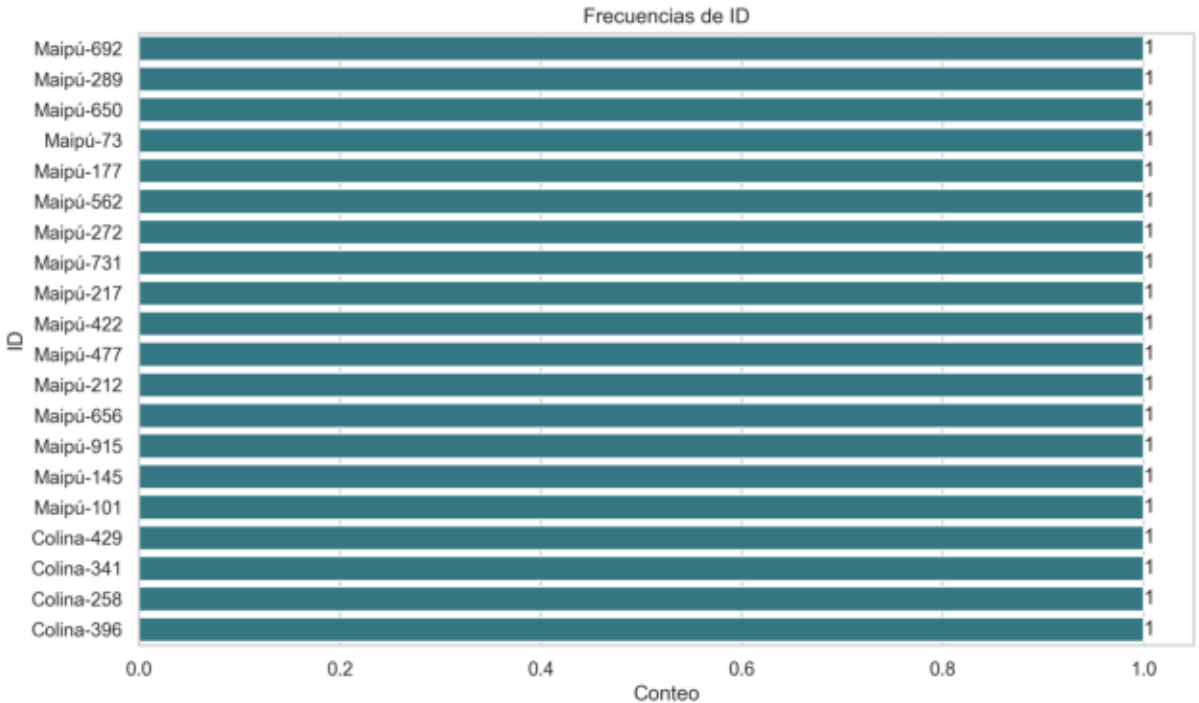
categoricas_barras_Calle.png



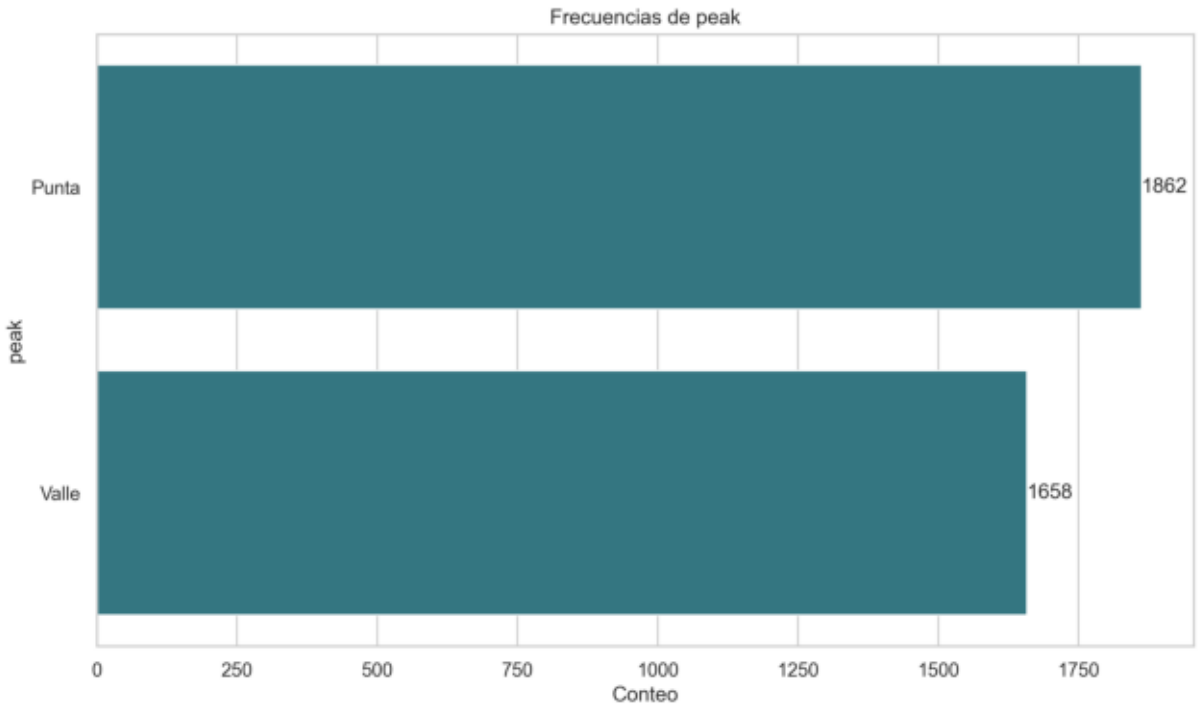
categoricas_barras_Comuna.png



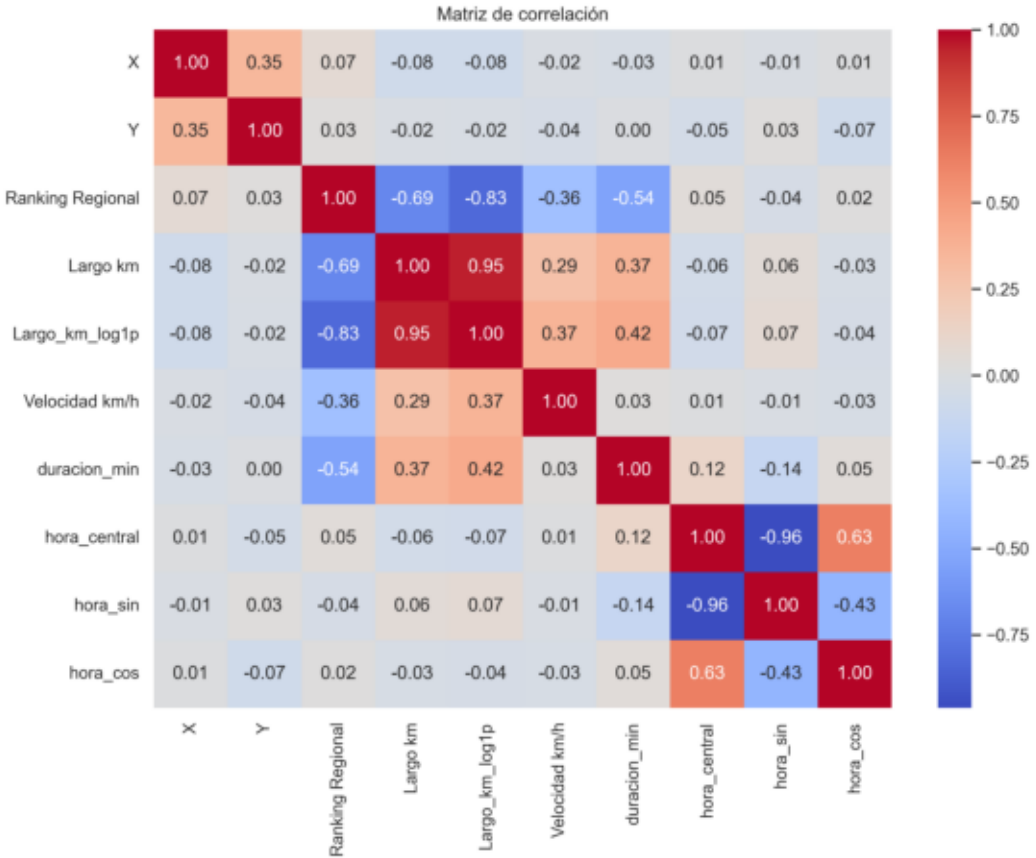
categoricas_barras_ID.png



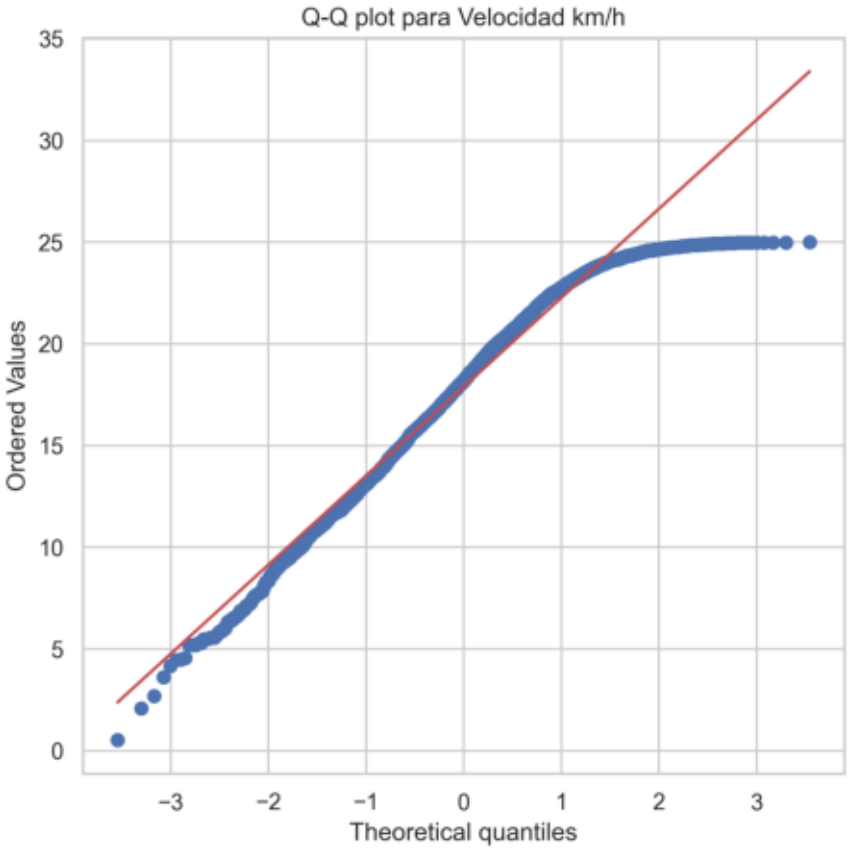
categoricas_barras_peak.png



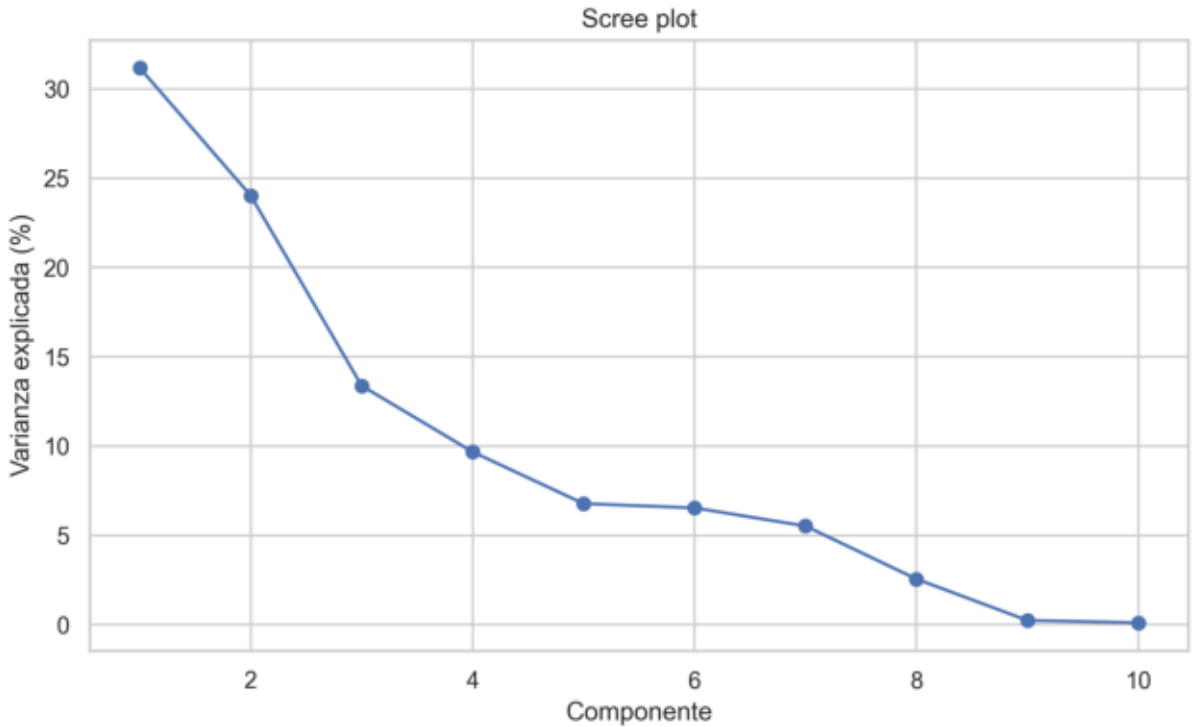
correlacion_heatmap.png



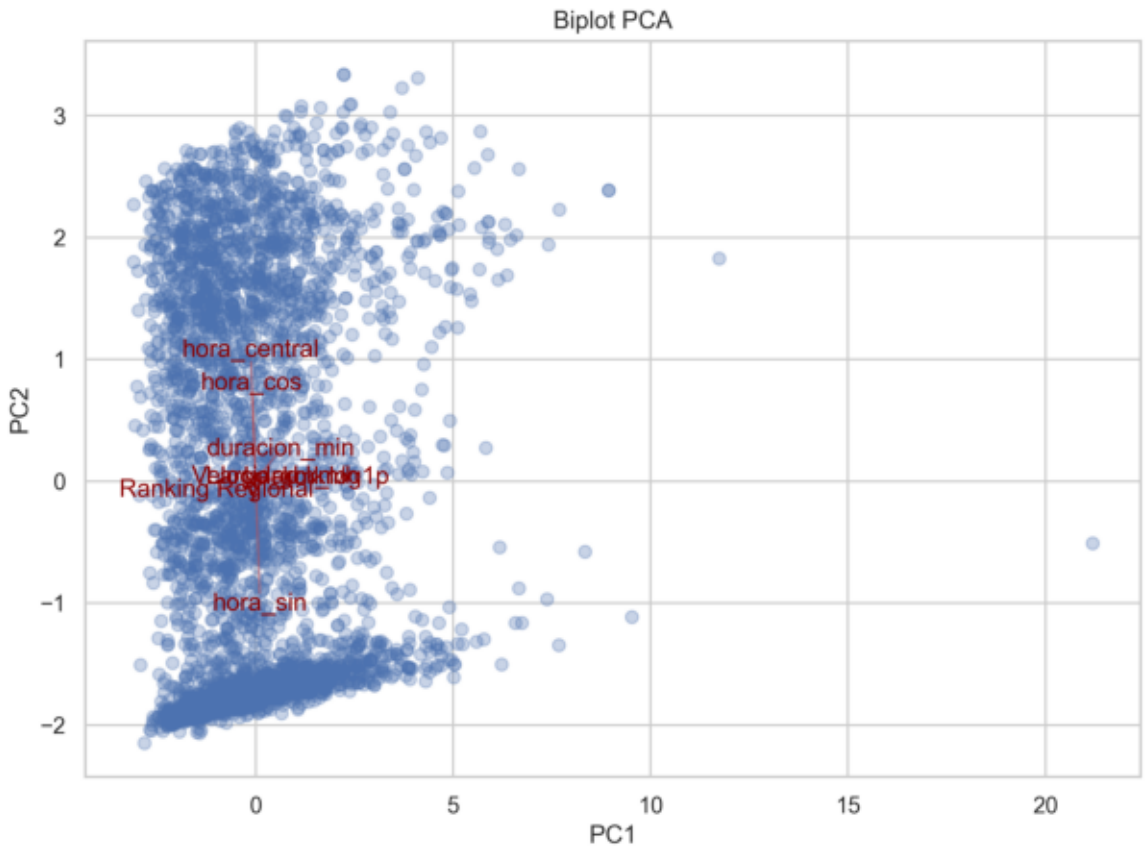
qqplot_Velocidad_kmh.png



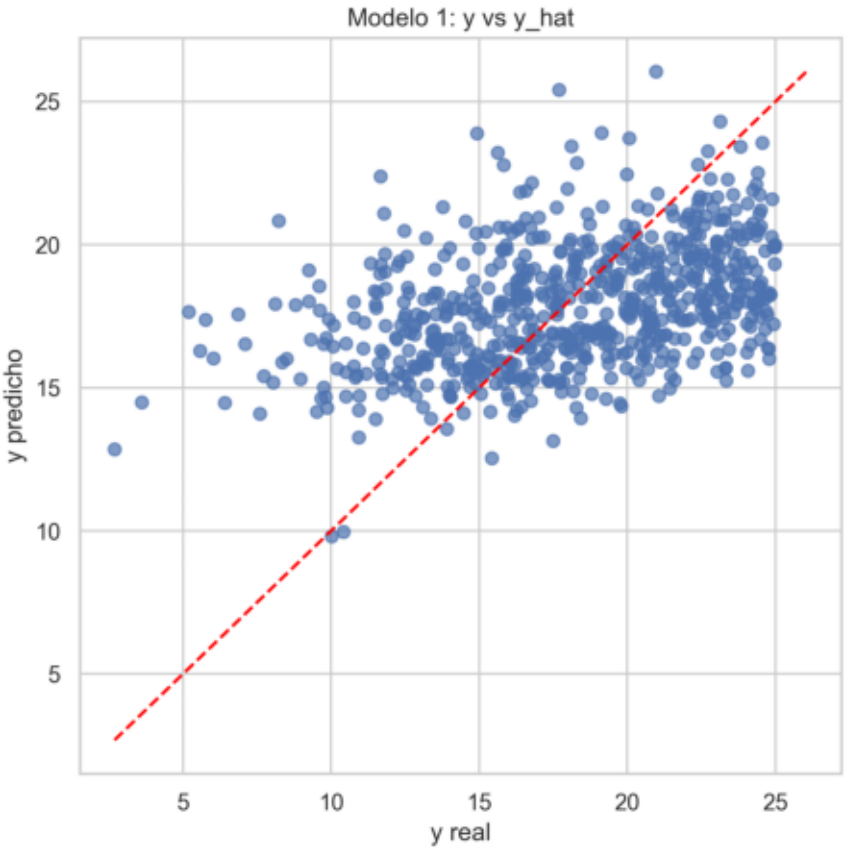
pca_varianza.png



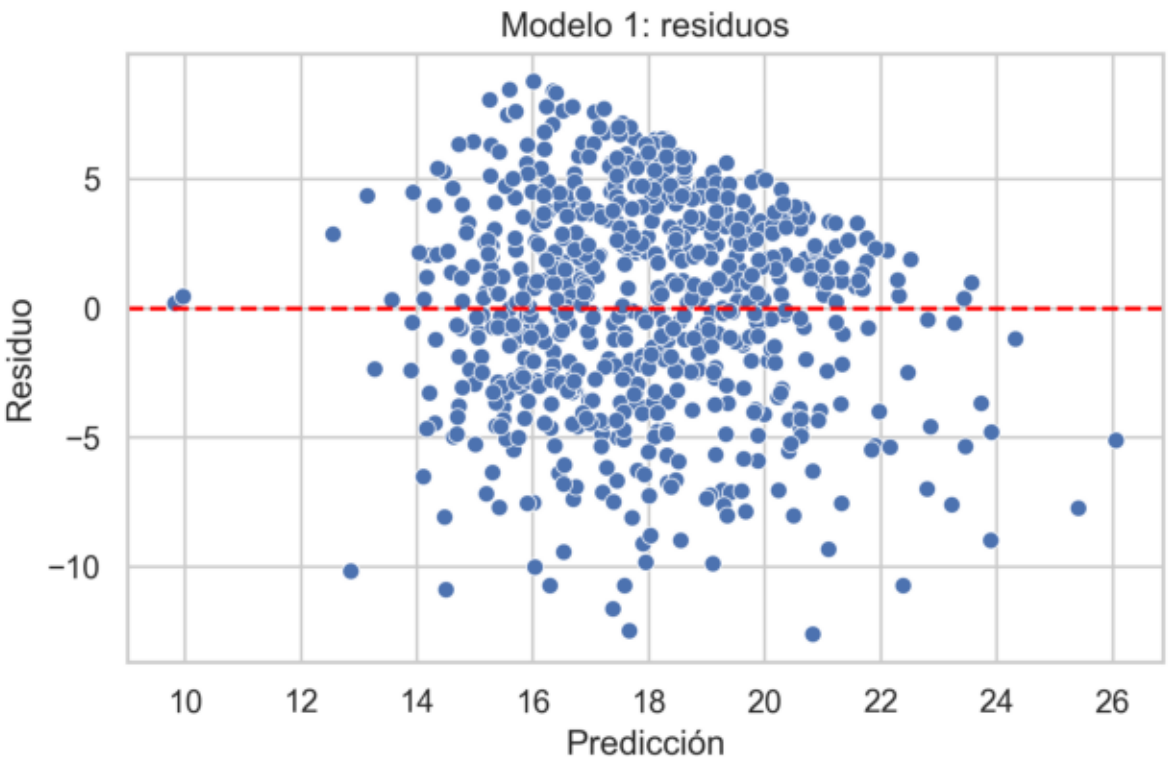
pca_biplot.png



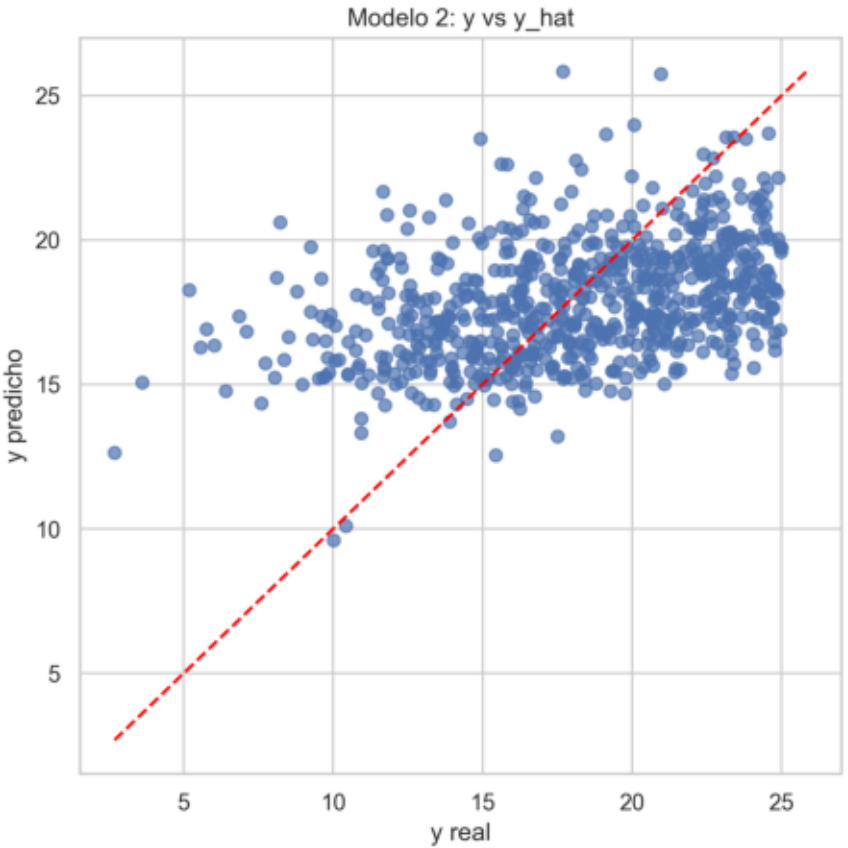
modelo_scatter_y_vs_yhat_m1.png



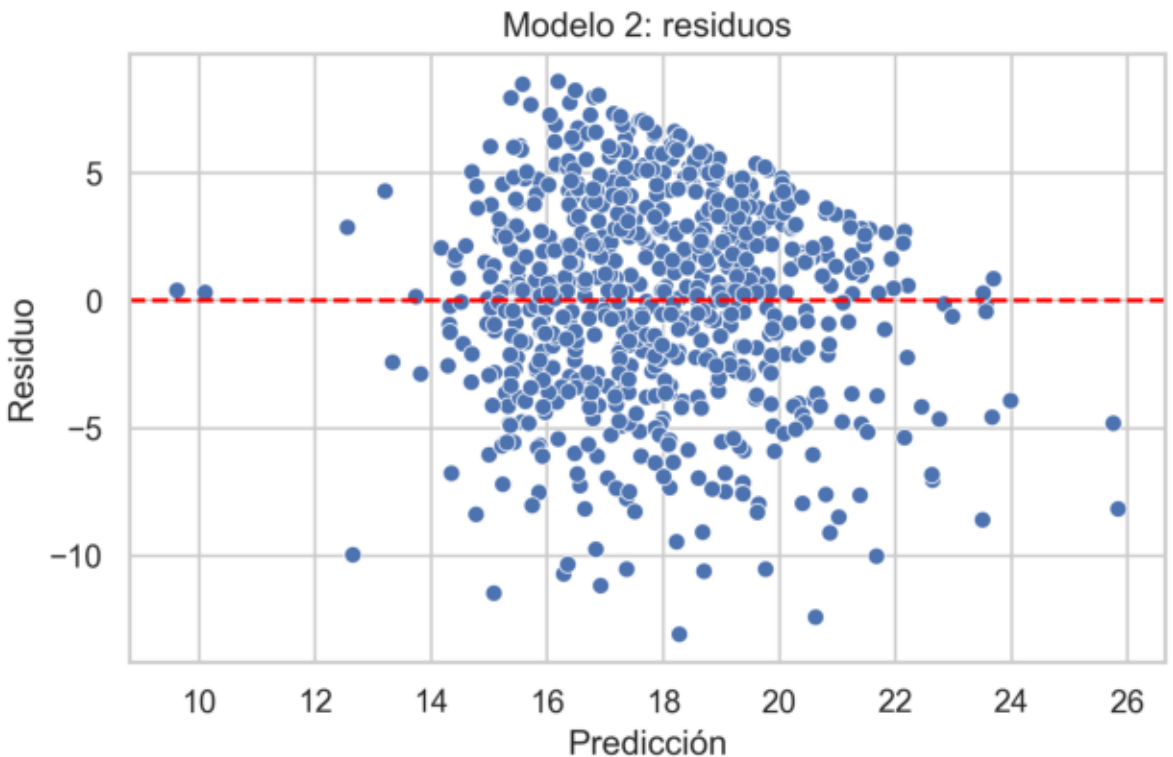
residuos_m1.png



modelo_scatter_y_vs_yhat_m2.png



residuos_m2.png



Conclusiones

Hallazgos:

- Modelo 1 resalta a Comuna_Pudahuel con coeficiente -2.956 (p=0.012).

- Modelo 1 resalta a Comuna_Recoleta con coeficiente -2.086 (p=0.074).

- Modelo 1 resalta a Comuna_Peñalolén con coeficiente -1.885 (p=0.066).

- Modelo 1 resalta a Comuna_Lampa con coeficiente -1.789 (p=0.263).

- Modelo 1 resalta a Comuna_Quinta Normal con coeficiente -1.774 (p=0.158).

- Modelo 2 identifica a Comuna_Pudahuel con coeficiente -2.213 (p=0.000).

- Modelo 2 identifica a Comuna_La Pintana con coeficiente 1.585 (p=0.027).

- Modelo 2 identifica a Ranking Regional con coeficiente -1.480 (p=0.000).

- Modelo 2 identifica a duracion_min con coeficiente -1.084 (p=0.000).

- Modelo 2 identifica a Comuna_Recoleta con coeficiente -0.971 (p=0.033).

- Correlación destacada entre hora_central y hora_sin: -0.96.

- Correlación destacada entre Largo km y Largo_km_log1p: 0.95.

- Correlación destacada entre Ranking Regional y Largo_km_log1p: -0.83.

- Correlación destacada entre Ranking Regional y Largo km: -0.69.

Referencias

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

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