PONTIFICIA UNIVERSIDAD CATÓLICA MADRE Y MAESTRA

Departamento de Ciencias e Ingeniería.

Escuela de Ciencias En Computación y Telecomunicaciones

MINERIA DE DATOS

"Proyecto Final"

Presentado por:

Nicol Ureña (2018-1669)

Junior Hernández (2018-0999)

Entregado a: Lisibonny Beato

Fecha de entrega: 8 de agosto del 2022

SANTIAGO DE LOS CABALLEROS, REPÚBLICA DOMINICANA

Índice

Business Understanding

Introducción

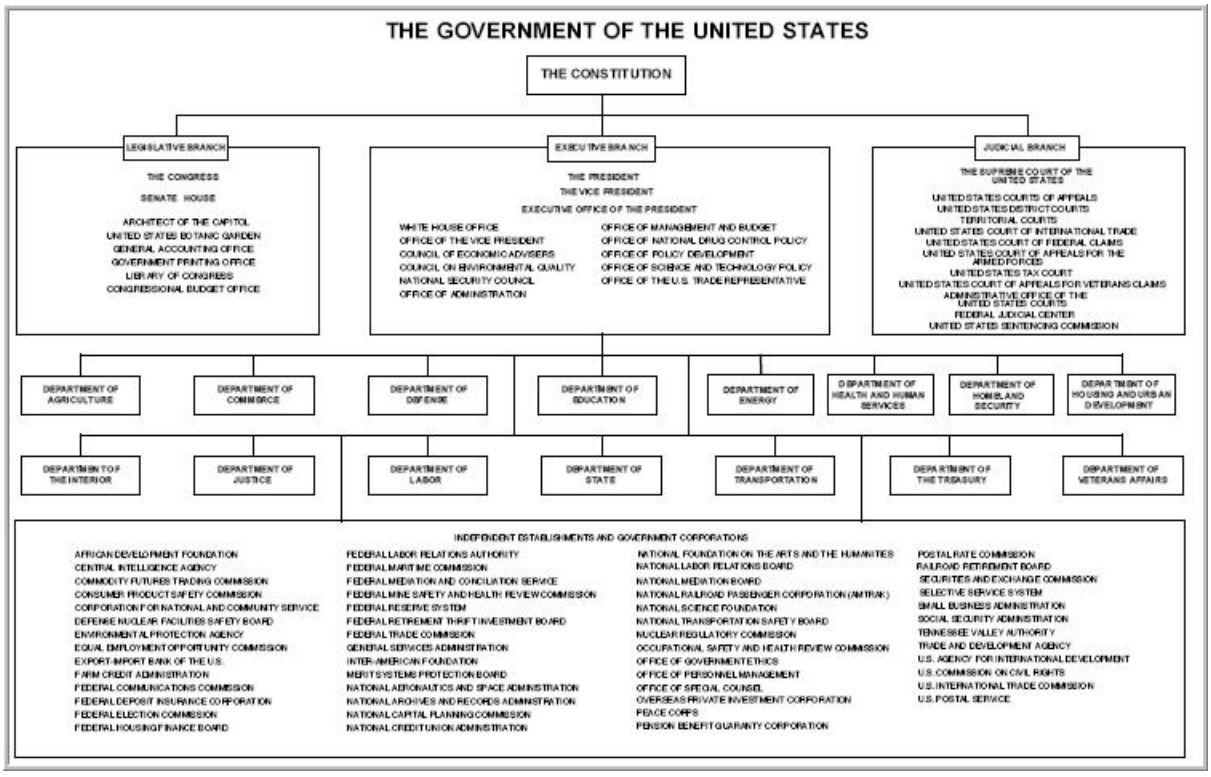
Una de las secretarías del DOT se ha comunicado con nosotros, dos ingenieros en sistemas de computación con conocimientos en minería de datos, donde el Departamento de Transportes de los Estados Unidos, que lleva un control detallado de los vuelos retrasados que se realizan dentro de su territorio, desea poder tener mejor control de la situación de retrasos en los vuelos, para poder así mejorar la satisfacción del cliente aéreo y la calidad del servicio de las líneas aéreas.

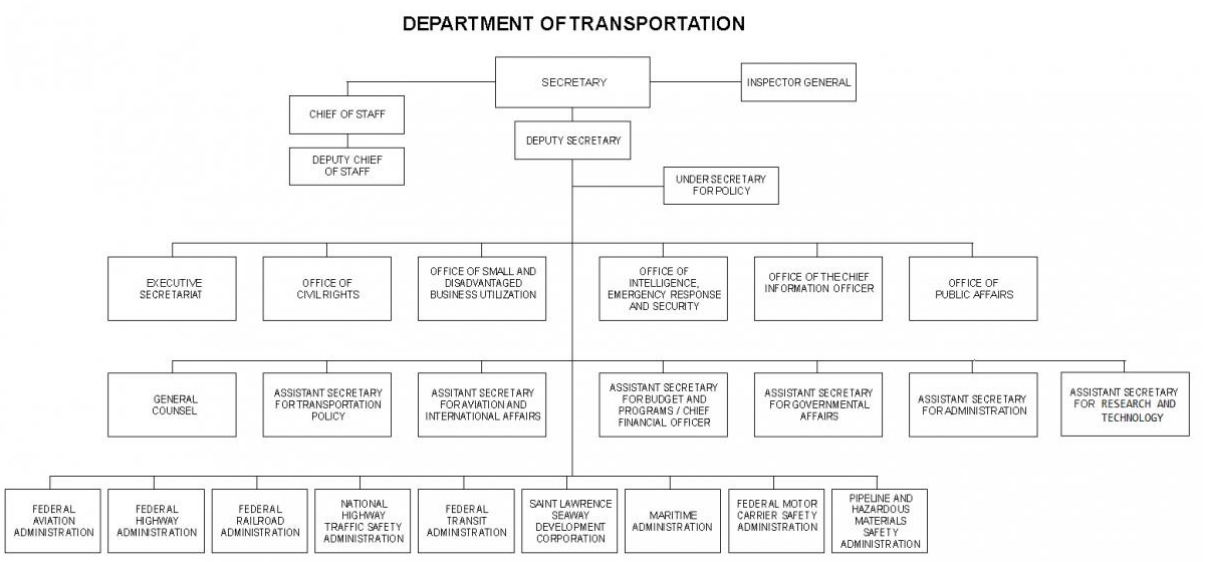
1.1 Determinar Objetivos del Negocio.

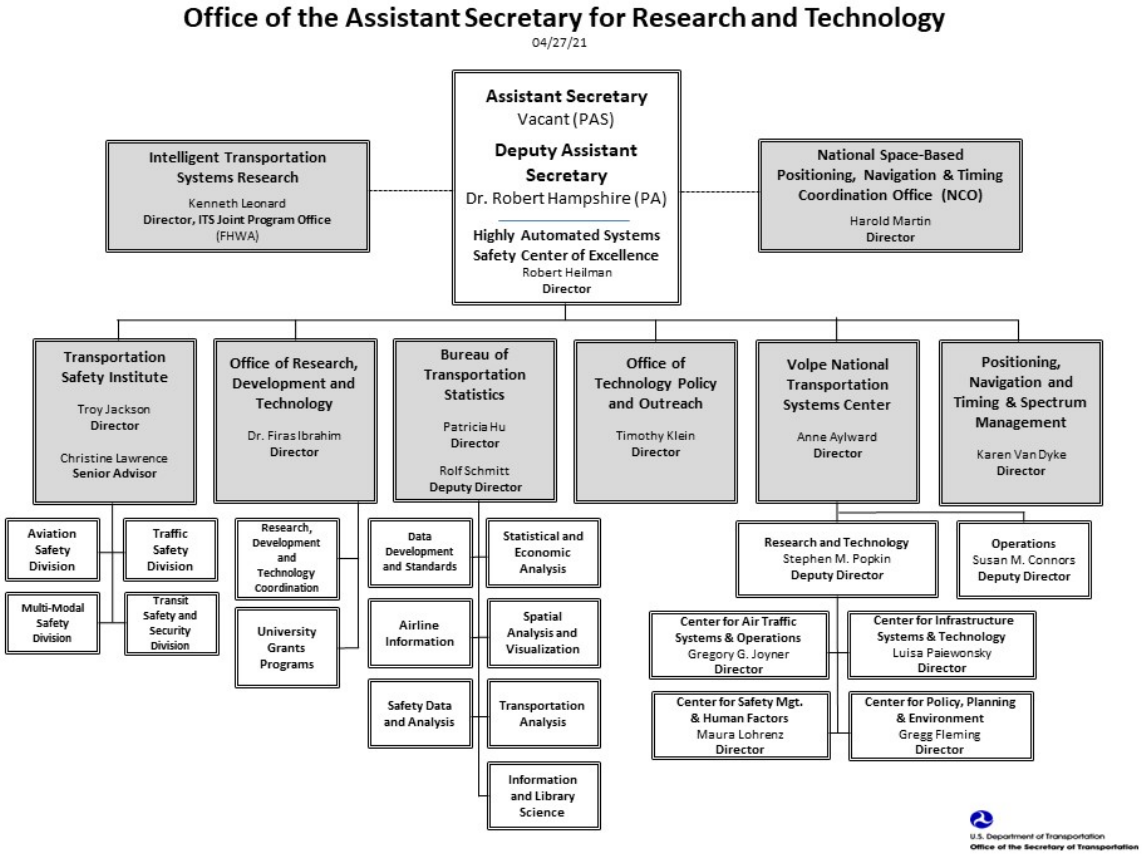
-Determinación de los objetivos comerciales

- Mejorar la satisfacción del cliente aéreo y la calidad del servicio de las líneas aéreas.
- Compilar los antecedentes comerciales.

- Determinar la estructura organizativa







-Describir el área del problema

El área encargada del problema es la oficina de la subsecretaría de investigación y tecnología.

- Describir la solución actual

El problema a solucionar es el cómo a través de los datos recopilados por la organización se puede tener mejor control de la situación de retrasos en los vuelos, para mejorar la satisfacción del cliente aéreo y la calidad del servicio de las líneas aéreas.

-Definición de objetivos comerciales

Mantener al público viajero seguro y protegido, aumentar su movilidad y hacer que nuestro sistema de transporte contribuya al crecimiento económico de la nación.

-Criterios de Éxito Empresarial

Recopilar estadísticas que son recogidas en distintos reportes detallando los vuelos retrasados que se realizan dentro de su territorio, a fin de obtener informaciones relevantes para la mejora de los servicios.

1.2 Evaluación de la situación

-Inventario de Recursos

La información disponible para trabajar con la problemática es la propuesta por Bureau of Transportation Statistics y Air Travel Consumer Reports del U.S. Department of Transportation, dichas fuentes probé las estadísticas y reportes acerca información sobre la calidad de los servicios prestados por las líneas aéreas, datos que van desde 2020 hasta el mes más actual del 2022, y la puntualidad de las aerolíneas y causas de los retrasos en los vuelos, uno de los grandes riesgos en que nos enfrentamos es el de estudiar y transformar cada uno de estos datos y encontrar si existe una correlación entre ellos para poder determinar si existe un factor el cual el departamento de transporte pueda controlar los retrasos de sus vuelos.

-Recursos del Hardware

- Investigación de recursos de hardware.

  Computadora: Máquina electrónica capaz de almacenar información y tratarla automáticamente mediante operaciones matemáticas y lógicas controladas por programas informáticos.

-Fuente de datos y conocimientos

-Identificar fuentes de datos y almacenes de conocimientos

La procedencia de los datos es del Bureau of Transportation Statistics y del U.S. Department of Transportation.

Las estadísticas y reportes de la puntualidad de las aerolíneas y causas de los retrasos en los vuelos del Bureau of Transportation Statistics son información acerca del porcentaje de la ocurrencia de una causa del problema.

El Air Travel Consumer Report del U.S. Department of Transportation es un producto mensual de la Oficina de Protección al Consumidor de Aviación del Departamento de Transporte. El informe está diseñado para ayudar a los consumidores con información sobre la calidad de los servicios prestados por las aerolíneas. El informe más reciente se emitió el 20 de julio de 2022.

El objetivo general de ambas estadísticas es buscar solución y que causas son las más regulares que afectan en el retraso de vuelos y en la inconformidad de los pasajeros.

-Fuentes de Personal

- Identificar recursos del personal

Computadores, Internet, energía eléctrica, plataformas para reuniones virtuales y aplicaciones para la minería de datos.

- Requisitos, suposiciones y restricciones

- Requerimientos: Los requisitos necesarios son de tener conocimientos plenos de que se encarga la empresa además de poseer datos que nos puedan proporcionar una idea de lo que se puede ir de lo general a lo específico de cómo se encuentra un aspecto dentro del control del retraso de sus vuelos para mejorar su calidad.

- Suposiciones: Suponemos que se puede asumir que la información a que se nos expone es una información verídica e inalterada.

- Restricciones: Se nos puede presentar una escasez de información o muestras con estos nos referimos a que no se encuentre de una manera “no limpia” de forma correcta.

-Riesgos y contingencias

- Identificar los riesgos

- Riesgo de que la información dada por el DOT no contenga una correlación entre los datos o que dichos estén incompletos.
- Riego de que el proyecto tome más tiempo de lo estimado.
- Riesgo que el proyecto no alcance el resultado que se espera.

-Identificar las contingencias

- Analizar los datos obtenidos haciendo uso de una de ETL para limpiar y transformar los datos presentados, seleccionando así cada uno de los que puedan tener correlación entre sí.
- Tomar medidas de planificación para trabajar en el proyecto lo más óptimo, rápido y analizado lo más posible.
- Ir tomando apuntes sobre qué es lo que se ha hecho hasta el momento y que es lo que falta, para establecer un camino a cómo llegar al final de lo esperado.

-Terminología

-Identificar terminología

DOT: por sus siglas en inglés Department of Transportation. Es un departamento del Gabinete federal del gobierno de los Estados Unidos encargado del transporte.

Tiempo aéreo. Las horas de vuelo de una aeronave calculadas desde el momento en que una aeronave deja el suelo hasta que toca el suelo al final de una etapa de vuelo.

Retraso en la llegada El retraso en la llegada es igual a la diferencia de la hora de llegada real menos la hora de llegada programada. Un vuelo se considera puntual cuando llega menos de 15 minutos después de la hora de llegada publicada.

ID de aeropuerto Un número de identificación asignado por el DOT de EE. UU. para identificar un aeropuerto único. Utilice este campo para el análisis de aeropuertos en un rango de años porque un aeropuerto puede cambiar su código de aeropuerto y los códigos de aeropuerto se pueden reutilizar.

Sistema Informático de Reservas CRS. Los CRS brindan información sobre los horarios de las aerolíneas, las tarifas y la disponibilidad de asientos a las agencias de viajes y permiten a los agentes reservar asientos y emitir boletos.

Estándares Federales de Procesamiento de Información FIPS. Por lo general, se refiere a un código asignado a cualquiera de una variedad de entidades geográficas (por ejemplo, condados, estados, áreas metropolitanas, etc.). Los códigos FIPS están destinados a simplificar la recopilación, el procesamiento y la difusión de datos y recursos del Gobierno Federal.

Retraso de vuelo: consiste en el retraso en la salida y/o llegada del vuelo respecto a las horas inicialmente programadas en su reserva.

Calidad: conjunto de propiedades inherentes a una cosa que permite caracterizarla y valorarla con respecto a las restantes de su especie.

Cliente: Persona que utiliza los servicios de un profesional o de una empresa, especialmente la que lo hace regularmente.

Minería de datos: proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados.

Costos y Beneficios

-Estimación de costos

Teniendo en cuenta que si este proyecto es llevado a cabo, no solo se mejorará la experiencia y satisfacción de los clientes, sino que también permitirá que una mayor cantidad de público se vea inclinado a elegir tomar vuelos en lugar de otras alternativas, resultando en una mejora significativa para clientes actuales y una fuente de promoción para nuevos clientes. Evaluando todas las ventajas anteriores, se estima que el costo estimado relativo a la recolección de datos es de 2,000 USD, mientras que por su lado el desarrollo de implementar una solución sería aproximadamente de 2,500 USD, resultando en un total de 5,000USD agregando los costos operativos.

1.3 Determinar objetivos de la Minería de Datos.

- Determinar los objetivos de la minería de datos.

Como objetivo principal de la minería de datos se pretende obtener un conjunto de clústeres con una seed de 10 que ayuden a determinar las causas comunes que influyen en los retrasos de vuelos aéreos, y de igual forma otros datos complementarios como las características específicas de los vuelos, y así predecir con mayor facilidad estas eventualidades.

Para cumplir los criterios de éxito del proyecto es necesario aplicar el algoritmo K-Means y obtener los clusters que cumplan con los parámetros establecidos.

1.4 Producir un plan de proyecto.

FASE	TIEMPO	MIEMBROS
BUSINESS UNDERSTANDING	5 días	Todos
DATA UNDERSTANDING	5 días	Todos
DATA PREPARATION	7 días	Todos
MODELING	4 días	Todos
EVALUATION	2 días	Todos
DEPLOYMENT	6 días	Todos

Data Understanding

Recolectar data inicial.

> Reporte de la colección inicial de datos.

January 2022 Air Travel Consumer Report

Regularly Scheduled Flights Canceled 5% or More

Conceptos	Descripción
Nombre del archivo	T_ONTIME_MARKETING(Enero)
Tamaño	216 MB
Formato	Excel
Descripción	Es un documento que contiene datos del mes de enero acerca de la información sobre los vuelos por los transportistas que comercializan para ellos mismos y, en algunos casos, para socios regionales de código compartido.
Cantidad de Registro	563737
Cantidad de Atributos	119

Conceptos	Descripción
Nombre del archivo	T_ONTIME_MARKETING(Febrero)
Tamaño	200 MB
Formato	Excel
Descripción	Es un documento que contiene datos del mes de febrero acerca de la información sobre los vuelos por los transportistas que comercializan para ellos mismos y, en algunos casos, para socios regionales de código compartido.
Cantidad de Registro	519952
Cantidad de Atributos	119

Conceptos	Descripción
Nombre del archivo	T_ONTIME_MARKETING(Marzo)
Tamaño	229 MB
Formato	Excel
Descripción	Es un documento que contiene datos del mes de marzo acerca de la información sobre los vuelos por los transportistas que comercializan para ellos mismos y, en algunos casos, para socios regionales de código compartido.

Cantidad de Registro	590542
Cantidad de Atributos	119

Conceptos	Descripción
Nombre del archivo	T_ONTIME_MARKETING(Abril)
Tamaño	225 MB
Formato	Excel
Descripción	Es un documento que contiene datos del mes de abril acerca de la información sobre los vuelos por los transportistas que comercializan para ellos mismos y, en algunos casos, para socios regionales de código compartido.
Cantidad de Registro	580290
Cantidad de Atributos	119

Descripción del documento original

	A	B	C	D	E	F	G	H	I	J	K
1	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_DATE	MKT_UNIQUE_CARRIER	BRANDED_CODE_SHARE	MKT_CARRIER_AIRLINE_ID	MKT_CARRIER	MKT_CARRIER_FL_NUM
2	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1581
3	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1582
4	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1582
5	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1583
6	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1584
7	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1584
8	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1585
9	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1586
10	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1587
11	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1587
12	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1588
13	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1589
14	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1589
15	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1590
16	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1591
17	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1591
18	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1592
19	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1593
20	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1594
21	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1595
22	2022	1	1	6	4	1/6/2022 0:00	DL	DL	19790	DL	1596

SCH_OP_UNIQUE_CARRIER	SCH_OP_CARRIER_AIRLINE_ID	SCH_OP_CARRIER	SCH_OP_CARRIER_FL_NUM	OP_UNIQUE_CARRIER	OP_CARRIER_AIRLINE_ID	OP_CARRIER	TAIL_NUM	OP_CARRIER_FL_NUM
				DL	19790	DL	N315DN	1581
				DL	19790	DL	N545US	1582
				DL	19790	DL	N545US	1582
				DL	19790	DL	N345NB	1583
				DL	19790	DL	N978AT	1584
				DL	19790	DL	N978AT	1584
				DL	19790	DL	N878DN	1585
				DL	19790	DL	N596NW	1586
				DL	19790	DL	N872DN	1587
				DL	19790	DL	N872DN	1587
				DL	19790	DL	N814DN	1588
				DL	19790	DL	N352DN	1589
				DL	19790	DL	N352DN	1589
				DL	19790	DL	N3769L	1590
				DL	19790	DL	N964AT	1591
				DL	19790	DL	N964AT	1591
				DL	19790	DL	N934AT	1592
				DL	19790	DL	N338DN	1593
				DL	19790	DL	N912DU	1594
				DL	19790	DL	N359NB	1595
				DL	19790	DL	N107DU	1596

OP_UNIQUE_CARRIER	OP_CARRIER_AIRLINE_ID	OP_CARRIER	TAIL_NUM	OP_CARRIER_FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN_AIRPORT_SEQ_ID	ORIGIN_CITY_MARKET_ID	ORIGIN	ORIGIN_CITY_NAME
DL	19790	DL	N315DN	1581	11697	1169706	32467	FLL	Fort Lauderdale, FL
DL	19790	DL	N545US	1582	10397	1039707	30397	ATL	Atlanta, GA
DL	19790	DL	N545US	1582	11697	1169706	32467	FLL	Fort Lauderdale, FL
DL	19790	DL	N345NB	1583	11697	1169706	32467	FLL	Fort Lauderdale, FL
DL	19790	DL	N978AT	1584	10397	1039707	30397	ATL	Atlanta, GA
DL	19790	DL	N978AT	1584	12448	1244807	32448	JAN	Jackson/Vicksburg, MS
DL	19790	DL	N878DN	1585	14524	1452401	34524	RIC	Richmond, VA
DL	19790	DL	N596NW	1586	13487	1348702	31650	MSP	Minneapolis, MN
DL	19790	DL	N872DN	1587	10397	1039707	30397	ATL	Atlanta, GA
DL	19790	DL	N872DN	1587	14492	1449202	34492	RDU	Raleigh/Durham, NC
DL	19790	DL	N814DN	1588	10693	1069302	30693	BNA	Nashville, TN
DL	19790	DL	N352DN	1589	10397	1039707	30397	ATL	Atlanta, GA
DL	19790	DL	N352DN	1589	12339	1233904	32337	IND	Indianapolis, IN
DL	19790	DL	N3769L	1590	12953	1295304	31703	LGA	New York, NY
DL	19790	DL	N964AT	1591	10397	1039707	30397	ATL	Atlanta, GA
DL	19790	DL	N964AT	1591	14685	1468502	34685	SAV	Savannah, GA
DL	19790	DL	N934AT	1592	15919	1591904	31834	XNA	Fayetteville, AR
DL	19790	DL	N338DN	1593	14683	1468305	33214	SAT	San Antonio, TX
DL	19790	DL	N912DU	1594	15304	1530402	33195	TPA	Tampa, FL
DL	19790	DL	N359NB	1595	14869	1486903	34614	SLC	Salt Lake City, UT
DL	19790	DL	N107DU	1596	14869	1486903	34614	SLC	Salt Lake City, UT

ORIGIN_STATE_ABR	ORIGIN_STATE_FIPS	ORIGIN_STATE_NM	ORIGIN_WAC	DEST_AIRPORT_ID	DEST_AIRPORT_SEQ_ID	DEST_CITY_MARKET_ID	DEST	DEST_CITY_NAME
FL	12	Florida	33	12953	1295304	31703	LGA	New York, NY
GA	13	Georgia	34	11697	1169706	32467	FLL	Fort Lauderdale, FL
FL	12	Florida	33	10397	1039707	30397	ATL	Atlanta, GA
FL	12	Florida	33	14492	1449202	34492	RDU	Raleigh/Durham, NC
GA	13	Georgia	34	12448	1244807	32448	JAN	Jackson/Vicksburg, MS
MS	28	Mississippi	53	10397	1039707	30397	ATL	Atlanta, GA
VA	51	Virginia	38	10397	1039707	30397	ATL	Atlanta, GA
MN	27	Minnesota	63	14635	1463502	31714	RSW	Fort Myers, FL
GA	13	Georgia	34	14492	1449202	34492	RDU	Raleigh/Durham, NC
NC	37	North Carolina	36	10397	1039707	30397	ATL	Atlanta, GA
TN	47	Tennessee	54	10397	1039707	30397	ATL	Atlanta, GA
GA	13	Georgia	34	12339	1233904	32337	IND	Indianapolis, IN
IN	18	Indiana	42	10397	1039707	30397	ATL	Atlanta, GA
NY	36	New York	22	11292	1129202	30325	DEN	Denver, CO
GA	13	Georgia	34	14685	1468502	34685	SAV	Savannah, GA
GA	13	Georgia	34	10397	1039707	30397	ATL	Atlanta, GA
AR	5	Arkansas	71	10397	1039707	30397	ATL	Atlanta, GA
TX	48	Texas	74	10397	1039707	30397	ATL	Atlanta, GA
FL	12	Florida	33	11433	1143302	31295	DTW	Detroit, MI
UT	49	Utah	87	11292	1129202	30325	DEN	Denver, CO
UT	49	Utah	87	11292	1129202	30325	DEN	Denver, CO
GA	13	Georgia	34	12339	1233904	32337	IND	Indianapolis, IN

DEST_STATE_ABR	DEST_STATE_FIPS	DEST_STATE_NM	DEST_WAC	CRS_DEP_TIME	DEP_TIME	DEP_DELAY	DEP_DELAY_NEW	DEP_DEL15	DEP_DELAY_GROUP	DEP_TIME_BLK
NY	36	New York	22	1126						1100-1159
FL	12	Florida	33	1631	1627	-4	0	0	-1	1600-1659
GA	13	Georgia	34	1931	1929	-2	0	0	-1	1900-1959
NC	37	North Carolina	36	1024	1019	-5	0	0	-1	1000-1059
MS	28	Mississippi	53	1117	1113	-4	0	0	-1	1100-1159
GA	13	Georgia	34	1237	1230	-7	0	0	-1	1200-1259
GA	13	Georgia	34	900	857	-3	0	0	-1	0900-0959
FL	12	Florida	33	1000	1018	18	18	1	1	1000-1059
NC	37	North Carolina	36	1414	1422	8	8	0	0	1400-1459
GA	13	Georgia	34	1643	1640	-3	0	0	-1	1600-1659
GA	13	Georgia	34	701	656	-5	0	0	-1	0700-0759
IN	18	Indiana	42	1235	1234	-1	0	0	-1	1200-1259
GA	13	Georgia	34	1511	1504	-7	0	0	-1	1500-1559
CO	8	Colorado	82	1826	1823	-3	0	0	-1	1800-1859
GA	13	Georgia	34	820	816	-4	0	0	-1	0800-0859
GA	13	Georgia	34	1022	1013	-9	0	0	-1	1000-1059
GA	13	Georgia	34	600	557	-3	0	0	-1	0600-0659
GA	13	Georgia	34	700	656	-4	0	0	-1	0700-0759
MI	26	Michigan	43	1815	1810	-5	0	0	-1	1800-1859
CO	8	Colorado	82	955	1046	51	51	1	3	0900-0959
CO	8	Colorado	82	2105	2101	-4	0	0	-1	2100-2159

TAXI_OUT	WHEELS_OFF	WHEELS_ON	TAXI_IN	CRS_ARR_TIME	ARR_TIME	ARR_DELAY	ARR_DELAY_NEW	ARR_DEL15	ARR_DELAY_GROUP	ARR_TIME_BLK	CANCELLED	CANCELLATION_CODE	DIVERTED	DUP
				1419						1400-1459	1	A	0	N
15	1642	1815	5	1821	1820	-1	0	0	-1	1800-1859	0		0	N
13	1942	2105	10	2127	2115	-12	0	0	-1	2100-2159	0		0	N
17	1036	1209	3	1227	1212	-15	0	0	-1	1200-1259	0		0	N
14	1127	1127	4	1142	1131	-11	0	0	-1	1100-1159	0		0	N
11	1241	1430	5	1458	1435	-23	0	0	-2	1400-1459	0		0	N
13	910	1030	8	1050	1038	-12	0	0	-1	1000-1059	0		0	N
15	1033	1431	4	1433	1435	2	2	0	0	1400-1459	0		0	N
11	1433	1528	3	1533	1531	-2	0	0	-1	1500-1559	0		0	N
12	1652	1759	4	1814	1803	-11	0	0	-1	1800-1859	0		0	N
18	714	853	6	910	859	-11	0	0	-1	0900-0959	0		0	N
11	1245	1349	6	1401	1355	-6	0	0	-1	1400-1459	0		0	N
11	1515	1624	6	1643	1630	-13	0	0	-1	1600-1659	0		0	N
17	1840	2036	11	2112	2047	-25	0	0	-2	2100-2159	0		0	N
16	832	909	3	922	912	-10	0	0	-1	0900-0959	0		0	N
11	1024	1108	7	1132	1115	-17	0	0	-2	1100-1159	0		0	N
12	609	825	5	903	830	-33	0	0	-2	0900-0959	0		0	N
12	708	955	27	1014	1022	8	8	0	0	1000-1059	0		0	N
11	1821	2026	6	2102	2032	-30	0	0	-2	2100-2159	0		0	N
14	1100	1156	18	1134	1214	40	40	1	2	1100-1159	0		0	N
16	2117	2210	24	2238	2234	-4	0	0	-1	2200-2259	0		0	N

CRS_ELAPSED_TIME	ACTUAL_ELAPSED_TIME	AIR_TIME	FLIGHTS	DISTANCE	DISTANCE_GROUP	CARRIER_DELAY	WEATHER_DELAY	NAS_DELAY	SECURITY_DELAY	LATE_AIRCRAFT_DELAY	FIRST_DEP_TIME
173			1	1076	5						
110	113	93	1	581	3						
116	106	83	1	581	3						
123	113	93	1	680	3						
85	78	60	1	341	2						
81	65	49	1	341	2						
110	101	80	1	481	2						
213	197	178	1	1416	6						
79	69	55	1	356	2						
91	83	67	1	356	2						
69	63	39	1	214	1						
86	81	64	1	432	2						
92	86	69	1	432	2						
286	264	236	1	1620	7						
62	56	37	1	214	1						
70	62	44	1	214	1						
123	93	76	1	589	3						
134	146	107	1	874	4						
167	142	125	1	983	4						
99	88	56	1	391	2	2	0	0	0	38	
93	93	53	1	391	2						

TOTAL_ADD_GTIME	LONGEST_ADD_GTIME	DIV_AIRPORT_LANDINGS	DIV_REACHED_DEST	DIV_ACTUAL_ELAPSED_TIME	DIV_ARR_DELAY	DIV_DISTANCE	DIV1_AIRPORT	DIV1_AIRPORT_ID	DIV1_AIRPORT_SEQ_ID
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							
		0							

DIV1_WHEELS_ON	DIV1_TOTAL_GTIME	DIV1_LONGEST_GTIME	DIV1_WHEELS_OFF	DIV1_TAIL_NUM	DIV2_AIRPORT	DIV2_AIRPORT_ID	DIV2_AIRPORT_SEQ_ID	DIV2_WHEELS_ON	DIV2_TOTAL_GTIME

DIV2_LONGEST_GTIME	DIV2_WHEELS_OFF	DIV2_TAIL_NUM	DIV3_AIRPORT	DIV3_AIRPORT_ID	DIV3_AIRPORT_SEQ_ID	DIV3_WHEELS_ON	DIV3_TOTAL_GTIME	DIV3_LONGEST_GTIME	DIV3_TAIL_NUM

DIV4_AIRPORT	DIV4_AIRPORT_ID	DIV4_AIRPORT_SEQ_ID	DIV4_WHEELS_ON	DIV4_TOTAL_GTIME	DIV4_LONGEST_GTIME	DIV4_WHEELS_OFF	DIV4_TAIL_NUM	DIV5_AIRPORT	DIV5_AIRPORT_ID

DIV5_AIRPORT_SEQ_ID	DIV5_WHEELS_ON	DIV5_TOTAL_GTIME	DIV5_LONGEST_GTIME	DIV5_WHEELS_OFF	DIV5_TAIL_NUM

La tabla muestra los datos de puntualidad para la red de transportistas de comercialización, si corresponde, el transportista de comercialización que es el transportista informador y los afiliados de código compartido del transportista de comercialización como grupo. Los transportistas que notifican que no comercializan vuelos se incluyen en el grupo de código compartido regional. La tabla muestra: datos de llegadas y salidas puntuales para vuelos nacionales sin escalas por mes y año, por red de comercialización, aerolínea de comercialización que informa y grupo regional de código compartido, por aeropuerto de origen y de destino. Incluye las horas de salida y llegada programadas y reales, los vuelos cancelados y desviados, las horas de salida y llegada del taxi, las causas de la demora y la cancelación, el tiempo de aire y la distancia sin escalas. Use Descargar para datos de vuelos individuales.

Reporte de descripción de datos

Nombre	**Tipo**	**Descripción**
Year	Num	Año
Time Period		
Quarter	Num	Quarter (1-4)
Month	Num	Mes

DayofMonth	Num	Dia del mes
DayOfWeek	Num	Dia de la semana
FlightDate	Nom	Fecha de vuelo (yyyymmdd)
Airline		
Marketing_Airline_Network	Nom	Código único de operador de comercialización. Cuando varios operadores han utilizado el mismo código, se utiliza un sufijo numérico para usuarios anteriores, por ejemplo, PA, PA(1), PA(2). Utilice este campo para el análisis en un rango de años.
Operated_or_Branded_Code_Share_Partners	Nom	Informes de socios de código compartido de marca o operados por transportistas

DOT_ID_Marketing_Airline	Num	Un número de identificación asignado por el DOT de EE. UU. para identificar una aerolínea (transportista) única. Una aerolínea (transportista) única se define como una que posee y reporta bajo el mismo certificado DOT independientemente de su código, nombre o compañía/corporación tenedora.
IATA_Code_Marketing_Airline	Nom	Código asignado por IATA y comúnmente utilizado para identificar a un transportista. Como el mismo código puede haber sido asignado a diferentes transportistas a lo largo del tiempo, el código no siempre es único. Para el análisis, utilice el Código Único de Transportista.
Flight_Number_Marketing_Airline	Num	Número de vuelo

Originally_Scheduled_Code_Share_Airline	Nom	Código único de operador operativo programado. Cuando varios operadores han utilizado el mismo código, se utiliza un sufijo numérico para usuarios anteriores, por ejemplo, PA, PA(1), PA(2). Utilice este campo para el análisis en un rango de años.
DOT_ID_Originally_Scheduled_Code_Share_Airline	Num	Un número de identificación asignado por el DOT de EE. UU. para identificar una aerolínea (transportista) única. Una aerolínea (transportista) única se define como una que posee y reporta bajo el mismo certificado DOT independientemente de su código, nombre o compañía/corporación tenedora.
IATA_Code_Originally_Scheduled_Code_Share_Airline	Nom	Código asignado por IATA y comúnmente utilizado para identificar a un transportista. Como el mismo código puede haber sido asignado a diferentes transportistas a lo largo del

		tiempo, el código no siempre es único. Para el análisis, utilice el Código Único de Transportista.
Flight_Num_Originally_Scheduled_Code_Share_Airline	Num	Número de vuelo
Operating_Airline	Nom	Código Único de Transportista. Cuando varios operadores han utilizado el mismo código, se utiliza un sufijo numérico para usuarios anteriores, por ejemplo, PA, PA(1), PA(2). Utilice este campo para el análisis en un rango de años.
DOT_ID_Operating_Airline	Num	Un número de identificación asignado por el DOT de EE. UU. para identificar una aerolínea (transportista) única. Una aerolínea (transportista) única se define como una que posee y reporta bajo el mismo certificado DOT independientemente de su código, nombre o compañía/corporación tenedora.

IATA_Code_Operating_Airline	Nom	Código asignado por IATA y comúnmente utilizado para identificar a un transportista. Como el mismo código puede haber sido asignado a diferentes transportistas a lo largo del tiempo, el código no siempre es único. Para el análisis, utilice el Código Único de Transportista.
Tail_Number	Nom	Número de cola
Flight_Number_Operating_Airline	Num	Número de vuelo
Origin		
OriginAirportID	Num	Aeropuerto de origen, ID del aeropuerto. Un número de identificación asignado por el DOT de EE. UU. para identificar un aeropuerto único. Utilice este campo para el análisis de aeropuertos en un rango de años

		porque un aeropuerto puede cambiar su código de aeropuerto y los códigos de aeropuerto se pueden reutilizar.
OriginAirportSeqID	Num	Aeropuerto de origen, ID de secuencia de aeropuerto. Un número de identificación asignado por el DOT de EE. UU. para identificar un aeropuerto único en un momento determinado. Los atributos del aeropuerto, como el nombre o las coordenadas del aeropuerto, pueden cambiar con el tiempo.
OriginCityMarketID	Num	Aeropuerto de origen, ID de mercado de la ciudad. City Market ID es un número de identificación asignado por el DOT de EE. UU. para identificar un mercado de ciudad. Utilice este campo para consolidar los aeropuertos que sirven al mismo mercado de la ciudad.

Origin	Nom	Aeropuerto de origen
OriginCityName	Nom	Aeropuerto de origen, nombre de la ciudad
OriginState	Nom	Aeropuerto de Origen, Código de Estado
OriginStateFips	Num	Aeropuerto Origen, Estado Fips
OriginStateName	Nom	Aeropuerto de Origen, Nombre del Estado
OriginWac	Num	Aeropuerto de origen, código de área mundial
Destination		
DestAirportID	Num	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a

		unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.
DestAirportSeqID	Num	Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
DestCityMarketID	Num	Aeropuerto de destino, ID de mercado de la ciudad. City Market ID es un número de identificación asignado por el DOT de EE. UU. para identificar un mercado de ciudad. Utilice este campo para consolidar los aeropuertos que sirven al mismo mercado de la ciudad.
Dest	Nom	Aeropuerto de destino

DestCityName	Nom	Aeropuerto de destino, nombre de la ciudad
DestState	Nom	Aeropuerto de destino, código de estado
DestStateFips	Num	Aeropuerto de Destino, Estado Fips
DestStateName	Nom	Aeropuerto de destino, nombre del estado
DestWac	Num	Aeropuerto de destino, código de área mundial
Departure Performance		
CRSDepTime	Num	Hora de salida del CRS (hora local: hhmm)

DepTime	Num	Hora de salida real (hora local: hhmm)
DepDelay	Num	Diferencia en minutos entre la hora de salida prevista y la real. Las salidas anticipadas muestran números negativos.
DepDelayMinutes	Num	Diferencia en minutos entre la hora de salida prevista y la real. Salidas anticipadas puestas a 0.
DepDel15	Num	Indicador de retraso de salida, 15 minutos o más (1=Sí)
DepartureDelayGroups	Num	Intervalos de retraso de salida, cada (15 minutos de <-15 a >180)
DepTimeBlk	Nom	Bloque de tiempo de salida de CRS, intervalos por hora

TaxiOut	Num	Tiempo de salida del taxi, en minutos
WheelsOff	Num	Hora de inactividad de las ruedas (hora local: hhmm)
Arrival Performance		
WheelsOn	Num	Wheels On Time (hora local: hhmm)
TaxiIn	Num	Taxi a tiempo, en minutos
CRSArrTime	Num	Hora de llegada del CRS (hora local: hhmm)
ArrTime	Num	Hora real de llegada (hora local: hhmm)
ArrDelay	Num	Diferencia en minutos entre la hora de llegada prevista y la real.

		Las llegadas anticipadas muestran números negativos.
ArrDelayMinutes	Num	Diferencia en minutos entre la hora de llegada prevista y la real. Las llegadas anticipadas se establecen en 0.
ArrDel15	Num	Indicador de retraso de llegada, 15 minutos o más (1=Sí)
ArrivalDelayGroups	Num	Intervalos de retraso de llegada, cada (15 minutos de <-15 a >180)
ArrTimeBlk	Nom	Bloque de tiempo de llegada de CRS, intervalos por hora
Cancellations and Diversions		
Cancelled	Num	Indicador de vuelo cancelado (1=Sí)

CancellationCode	Nom	Especifica el motivo de la cancelación
Diverted	Num	Indicador de vuelo desviado (1=Sí)
Duplicate	Nom	Bandera duplicada marcada Y si el vuelo se intercambia según los datos del Formulario-3A
Flight Summaries		
CRSElapsedTime	Num	CRS Tiempo de vuelo transcurrido, en minutos
ActualElapsedTime	Num	Tiempo de vuelo transcurrido, en minutos
AirTime	Num	Tiempo de vuelo, en minutos

Flights	Num	Número de vuelos
Distance	Num	Distancia entre aeropuertos (millas)
DistanceGroup	Num	Intervalos de distancia, cada 250 millas, por segmento de vuelo
Cause of Delay (Data starts 6/2003)		
CarrierDelay	Num	Retraso del operador, en minutos
WeatherDelay	Num	Tiempo de retraso, en minutos
NASDelay	Num	Demora del Sistema Aéreo Nacional, en Minutos
SecurityDelay	Num	Retraso de seguridad, en minutos

LateAircraftDelay	Num	Retraso de aeronave tardía, en minutos
Gate Return Information at Origin Airport (Data starts 10/2008)		
FirstDepTime	Num	Hora de salida de la primera puerta en el aeropuerto de origen
TotalAddGTime	Num	Tiempo total en tierra fuera de la puerta para regreso a la puerta o vuelo cancelado
LongestAddGTime	Num	Mayor tiempo fuera de la puerta de embarque para regreso a la puerta o vuelo cancelado
Diverted Airport Information (Data starts 10/2008)		
DivAirportLandings	Num	Número de aterrizajes en aeropuertos desviados

DivReachedDest	Num	Vuelo desviado que llega al indicador de destino programado (1=Sí)
DivActualElapsedTime	Num	Tiempo transcurrido del vuelo desviado que llega al destino programado, en minutos. La columna ActualElapsedTime permanece NULL para todos los vuelos desviados.
DivArrDelay	Num	Diferencia en minutos entre la hora de llegada prevista y la hora real de un vuelo desviado que llega al destino previsto. La columna ArrDelay permanece NULL para todos los vuelos desviados.
DivDistance	Num	Distancia entre el destino programado y el aeropuerto final desviado (millas). El valor será 0 para el vuelo desviado que llega al destino programado.

Div1Airport	Nom	Código de aeropuerto desviado1
Div1AirportID	Num	Identificación del aeropuerto del aeropuerto desviado 1. La identificación del aeropuerto es una clave única para un aeropuerto
Div1AirportSeqID	Num	ID de secuencia de aeropuerto de aeropuerto desviado 1. Clave única para información específica de tiempo para un aeropuerto
Div1WheelsOn	Num	Wheels On Time (hora local: hh mm) en código de aeropuerto desviado 1
Div1TotalGTime	Num	Tiempo total en tierra lejos de la puerta en el código de aeropuerto desviado1

Div1LongestGTime	Num	Mayor tiempo en tierra lejos de la puerta en el código de aeropuerto desviado1
Div1WheelsOff	Num	Hora de inactividad de ruedas (hora local: hh mm) en código de aeropuerto desviado 1
Div1TailNum	Nom	Número de cola de la aeronave para código de aeropuerto desviado1
Div2Airport	Nom	Código de aeropuerto desviado2
Div2AirportID	Num	Identificación del aeropuerto del aeropuerto desviado 2. La identificación del aeropuerto es una clave única para un aeropuerto
Div2AirportSeqID	Num	ID de secuencia de aeropuerto del aeropuerto desviado 2. Clave

		única para información específica de tiempo para un aeropuerto
Div2WheelsOn	Num	Wheels On Time (hora local: hh mm) en código de aeropuerto desviado 2
Div2TotalGTime	Num	Tiempo total en tierra lejos de la puerta en el código de aeropuerto desviado2
Div2LongestGTime	Num	Mayor tiempo en tierra lejos de la puerta en el código de aeropuerto desviado2
Div2WheelsOff	Num	Hora de inactividad de ruedas (hora local: hh mm) en código de aeropuerto desviado 2
Div2TailNum	Nom	Número de cola de la aeronave para código de aeropuerto desviado2

Div3Airport	**String**	Código de aeropuerto desviado3
Div3AirportID	**String**	Identificación del aeropuerto del aeropuerto desviado 3. La identificación del aeropuerto es una clave única para un aeropuerto
Div3AirportSeqID	**String**	ID de secuencia de aeropuerto del aeropuerto desviado 3. Clave única para información específica de tiempo para un aeropuerto
Div3WheelsOn	**String**	Wheels On Time (local time: hhmm) at Diverted Airport Code3
Div3TotalGTime	**String**	Tiempo total en tierra lejos de la puerta en el código de aeropuerto desviado3

Div3LongestGTime	**String**	Mayor tiempo en tierra lejos de la puerta en el aeropuerto desviado Code3
Div3WheelsOff	**String**	Hora de inactividad de las ruedas (hora local: hhmm) en el código de aeropuerto desviado3
Div3TailNum	**String**	Número de cola de la aeronave para código de aeropuerto desviado3
Div4Airport	**String**	Código de aeropuerto desviado4
Div4AirportID	**String**	Identificación del aeropuerto del aeropuerto desviado 4. La identificación del aeropuerto es una clave única para un aeropuerto
Div4AirportSeqID	**String**	ID de secuencia de aeropuerto del aeropuerto desviado 4. Clave

		única para información específica de tiempo para un aeropuerto
Div4WheelsOn	**String**	Wheels On Time (hora local: hh mm) en código de aeropuerto desviado 4
Div4TotalGTime	**String**	Tiempo total en tierra lejos de la puerta en el código de aeropuerto desviado4
Div4LongestGTime	**String**	Mayor tiempo en tierra lejos de la puerta en el aeropuerto desviado Code4
Div4WheelsOff	**String**	Wheels Off Time (local time: hhmm) at Diverted Airport Code4
Div4TailNum	**String**	Número de cola de la aeronave para código de aeropuerto desviado4

Div5Airport	**String**	Código de aeropuerto desviado5
Div5AirportID	**String**	Identificación del aeropuerto del aeropuerto desviado 5. La identificación del aeropuerto es una clave única para un aeropuerto
Div5AirportSeqID	**String**	ID de secuencia de aeropuerto del aeropuerto desviado 5. Clave única para información específica de tiempo para un aeropuerto
Div5WheelsOn	**String**	Wheels On Time (hora local: hh mm) en código de aeropuerto desviado 5
Div5TotalGTime	**String**	Tiempo total en tierra lejos de la puerta en el código de aeropuerto desviado 5

Div5LongestGTime	**String**	Mayor tiempo en tierra lejos de la puerta en el código de aeropuerto desviado 5
Div5WheelsOff	**String**	Hora de inactividad de ruedas (hora local: hh mm) en código de aeropuerto desviado 5
Div5TailNum	**String**	Número de cola de la aeronave para código de aeropuerto desviado5

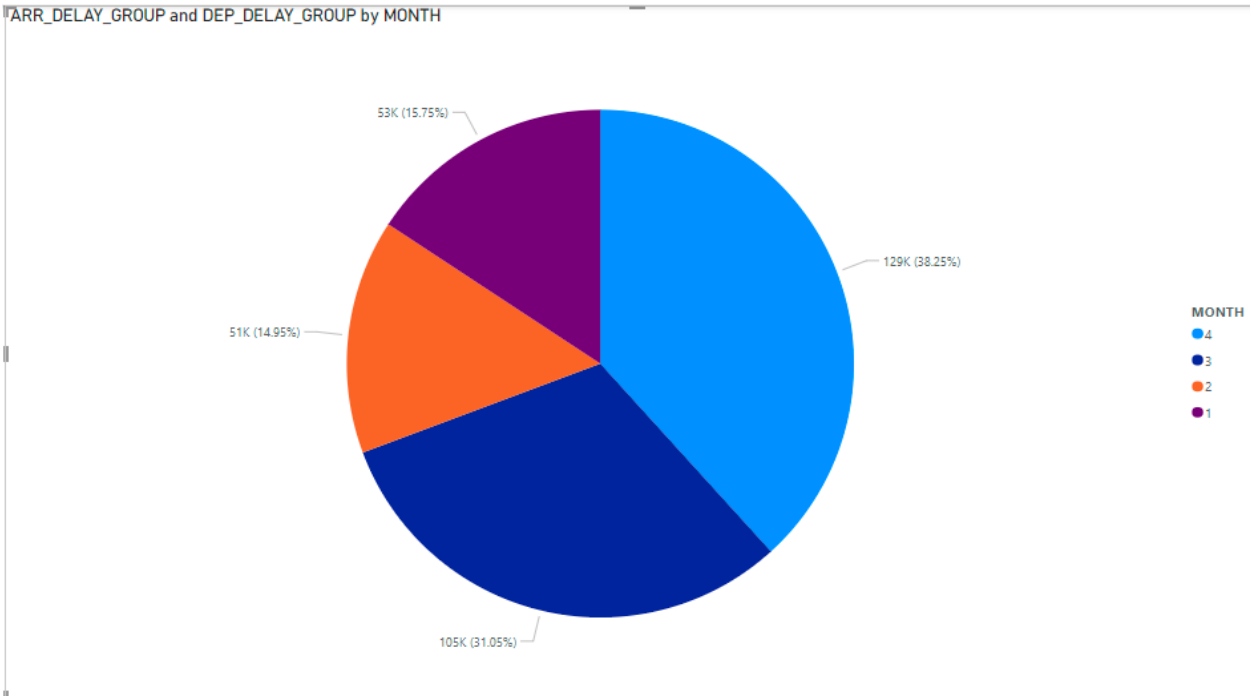
2.3 Exploración de la data.

De entre los atributos presentados, se observa que particularmente hay presentes ciertos atributos que son interesantes por el impacto que conllevan con los objetivos del negocio.

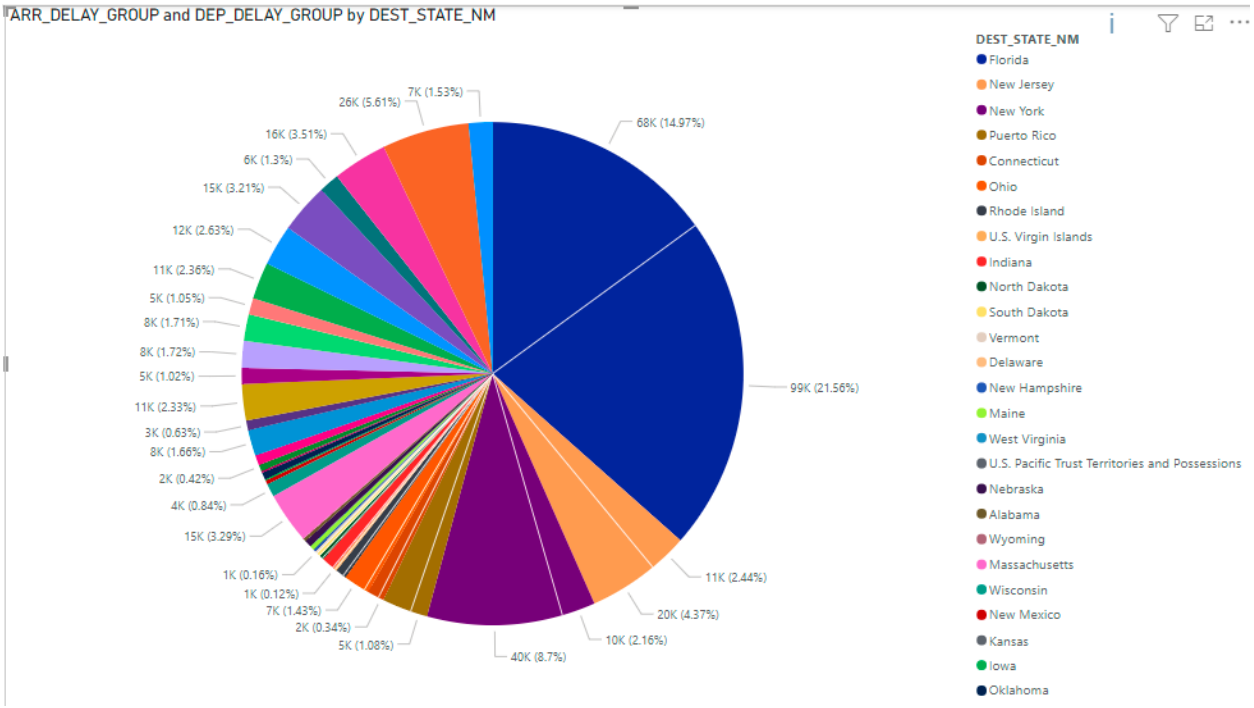
En primer lugar, unos datos esenciales son los referentes a DEP_DELAY_GROUP ARR_DELAY_GROUP, en donde se reflejan los grupos de retraso a los que pertenecen los vuelos. Este dato es el que permite que se registre la tardanza, y según aumenten o disminuyan los grupos que representan los minutos de tardanza, se podría visualizar una mejora (aunque también su contraparte) y evaluar que factores podrían influir en estos valores, a fin de tomar medidas que se acerquen a cumplir los objetivos propuestos. También caben destacar los atributos referentes a

los atrasos de carrier, clima, seguridad, nas y aircraft, que arrojan mucho que decir en cuanto a posibles análisis.

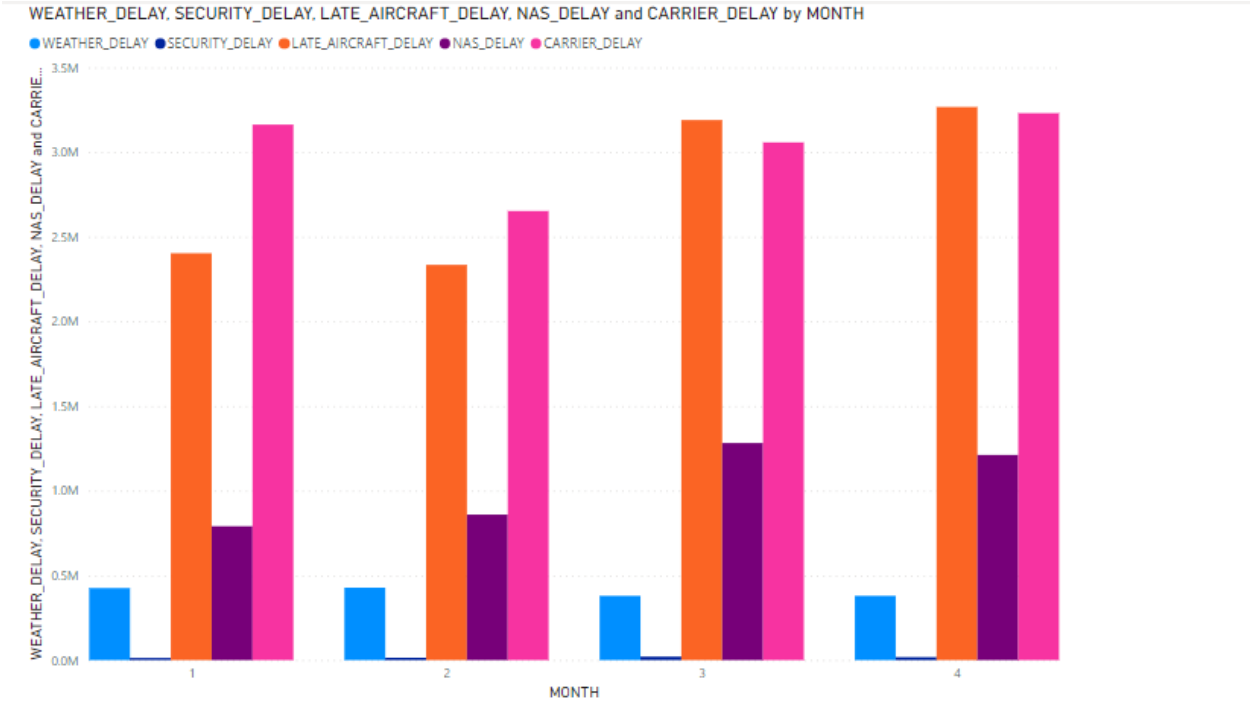
Por lo mencionado con anterioridad, se optó por graficar mediante la herramienta PowerBi los atributos de DEP_DELAY_GROUP ARR_DELAY_GROUP, es decir, el retraso de salida y retraso de llegada de los vuelos.



Aquí se observa por mes la cantidad de retrasos de salida y llegada de los vuelos, tomando en cuenta los meses de enero-abril del 2022. Se evidencia que la mayoría de retrasos ocurren en abril, seguido por marzo, enero, y febrero.



Aquí se observa por estado la cantidad de retrasos de salida y llegada de los vuelos, tomados en los meses de enero-abril del 2022. Se evidencia que la mayoría de retrasos ocurren en Florida, con una significativa diferencia en cuanto a cantidad si se compara con otros estados.



En este gráfico se presentan los 5 atributos de causas de los retrasos de vuelos, a comparación con los 4 meses empleados. En el mismo se puede apreciar los bajos que son los retrasos de seguridad y de clima en comparación con los retrasos de carrier o de avión tardío. Por otro lado, los delay de nas se mantienen en un término medio. No se observa mucha variabilidad entre los retrasos a través de los meses.

Preparación de los Datos

3.1 Selección de los Datos

En el dataset empleado, el referente a los meses de enero-abril de 2022 titulado T_ONTIME_MARKETING, se presenta un total de 119 atributos, en donde no todos eran pertinentes para los objetivos en cuestión, por lo que, podemos evidenciar que la lista de atributos se redujo significativamente.

En un primer lugar, los atributos que se mantuvieron fueron:

1. MONTH
2. DAY_OF_MONTH
3. OP_CARRIER
4. ORIGIN
5. ORIGIN_STATE_NM
6. DEST
7. DEST_STATE_NM
8. DEP_DELAY_GROUP
9. ARR_DELAY_GROUP
10. DISTANCE_GROUP
11. CARRIER_DELAY
12. WEATHER_DELAY
13. NAS_DELAY
14. SECURITY_DELAY
15. LATE_AIRCRAFT_DELAY

Como es observable, se emplearán un total de 15 atributos. En términos generales, los atributos que se eliminaron fueron aquellos innecesarios para el cumplimiento de los objetivos. Sin embargo, hubieron datos que presentaban características adicionales que los convertían en atributos completamente indebidos. Algunas de estas características fueron:

- Atributos sin datos, por lo que bajo esta limitante se eliminaron todas las columnas que no contenían ningún dato en ellas (como en el caso de DivReachedDest, etc.).

- Atributos repetidos. Debido a que eran datos que siempre se duplicaban, es decir, que sin excepción eran iguales a otro atributo, se prescindió de

MKT_UNIQUE_CARRIER (Igual a MKT _CARRIER)

OP_UNIQUE_CARRIER (Igual a OP _CARRIER)

OP_CARRIER_FL_NUM (Igual a MKT_CARRIER_FL_NUM)

-Los siguientes se descartaron debido a que se trata de abreviaciones de estados, que ya estaban incluidas en otro atributo.

ORIGIN_STATE_ABR (incluido en ORIGIN_CITY_NAME)

DEST_STATE_ABR (incluido en DEST_CITY_NAME)

-Atributos invariables. (FLIGHTS(invariable, 1 siempre), DUP (siempre N), Diverted).

- CANCELLED, no se encontró información de su significado A,B,C.

Entre otras condiciones.

3.2 Limpiar los datos.

En esta parte se consideró la calidad de los datos para su posterior análisis. De los atributos seleccionados, un total de 15, solo 5 atributos particulares, los cuales son CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY y LATE_AIRCRAFT_DELAY, fueron mantenidos como numéricos, lo demás se cambiaron o mantuvieron como nominales, de ahí en más, no se encontró ruido o atributos con instancias inadecuadas.

3.3 Construcción de los datos.

Atributos derivados

Las columnas de los grupos son atributos derivados que ya se encuentran en el dataset, fueron calculados por el departamento de transporte, estableciendo intervalos de tiempo para nivelarlos siendo 1 un intervalo de 15 o 30 minutos.

Filas generadas

Las filas tienen relación de acuerdo al mes. Por lo que los datos no numéricos y numéricos se pueden relacionar entre sí, debido a que fueron los últimos meses en donde se realizaron vuelos.

3.4. Integración de los datos

Se tiene en cuenta que los dataset recopilados tienen sus atributos iguales.

En los datasets se observa que los atributos se refieren a lo mismo, pero en meses diferentes. En donde envés de tener 4 dataset diferentes y evaluarlos por separados se unifican formando 1 solo. Y además en donde abundaba la sobre explicación de atributos siendo por ejemplo estableciendo una distancia para luego agruparla en un atributo donde explica los intervalos de cada una.

3.5 Format data

En el caso del formateo de los datos, nuestra herramienta a emplear, Weka, recibió un archivo en su formato de descarga de origen, .csv, aplicando filtros que permitieran a todos excepto los últimos 5, referentes al tipo de retraso, a ser de tipo nominal. Otro filtro aplicado fue el no supervisado de las instancias, en donde se aplicó el RemoveWithValues > matchMissingValues=True, es decir, eliminar las instancias que presentaban valores faltantes. El dataset con sus atributos en sus formatos correspondientes terminó siendo uno .arff gracias a la herramienta que realizó la conversión de tipos, y aunque en las partes donde hay valores que estén vacíos se escribe que es "?" para que Weka los ignore al procesarlos, al final estas instancias fueron eliminadas para cumplir mejor el objetivo planteado.

Modelado

4.1 Seleccionar la técnica de modelado.

En cuanto a la técnica de modelado empleada, hay que tomar en cuenta varias informaciones pertinentes. Utilizando la herramienta Weka en su versión 3.9.6, se plante utilizar la técnica de clustering, empleando el afamado algoritmo K-Means.

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster, por lo que emplea la distancia euclidiana.

Gracias a que este algoritmo permite utilizar tanto datos numéricos como nominales, se pudo implementar sin dificultad alguna. Se tiene en cuenta que para K-Means será necesario correrlo con diferentes tamaños de clúster, así a medida que se incrementa la cantidad de clústeres será posible ver en cuales permanecen más elementos y cuales tienen menos elementos. Para así ver en donde las variedades discrepan en lo que se refiere a sus atributos de los vuelos.

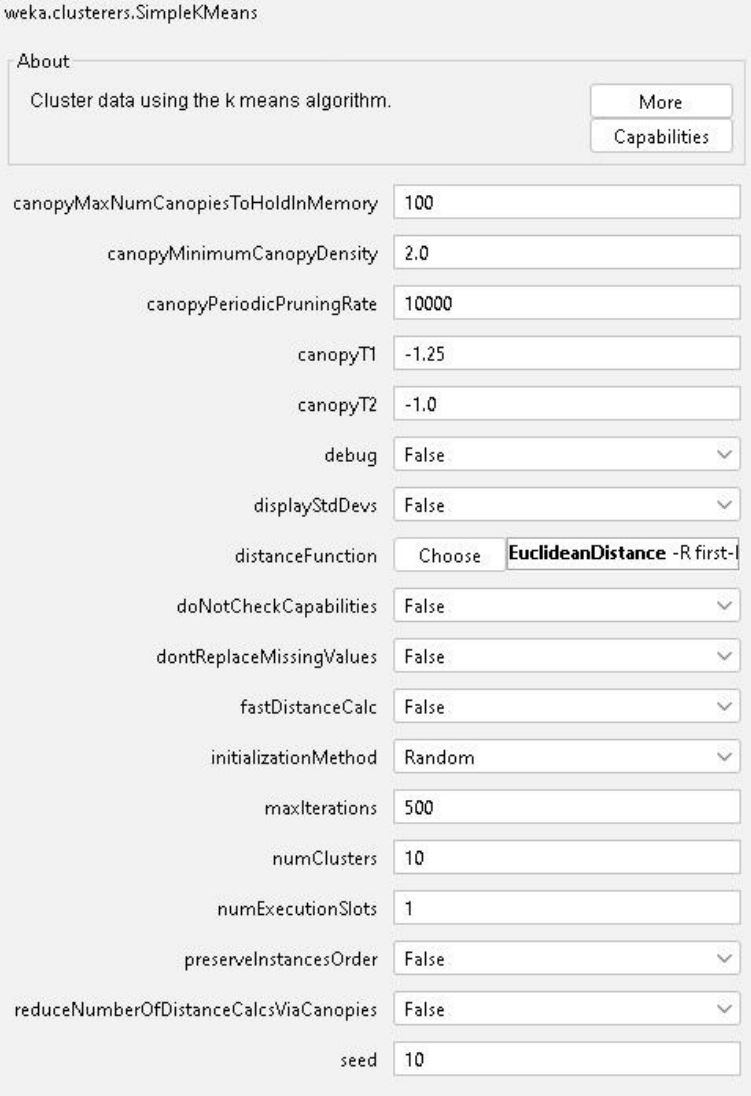
4.2 Generar test de diseño.

En el algoritmo K-Means, tenemos las siguiente fases para generar el modelo a partir de las pruebas:

- Como este algoritmo depende del número de clústeres que se le asigne, se plantea hacer varias pruebas utilizando diferentes números de clúster, empezando desde 2 hasta finalizar en 10.
- Los datos a utilizar serán los que se prepararon en la tercera fase de este proyecto, Data Preparation (Preparación de los datos)

4.3 Construir el Modelo.

En la construcción del modelo se tomaron en cuenta los siguientes parámetros:



Estos parámetros se utilizaron debido a que luego de realizar distintas pruebas, se obtuvieron los siguientes resultados para n clústeres asignados al algoritmo k-means:

K-Means (n =2)

```
Cluster 0: 4,12,UA,SFO,California,IAD,Virginia,3,3,10,45,0,2,0,0
Cluster 1: 4,10,QX,SEA,Washington,BLI,Washington,1,1,1,2,0,0,0,21

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                        Cluster#
Attribute                                  Full Data          0          1
                                          (406372.0) (145182.0) (261190.0)
=============================================================================
MONTH                                             4          3          4
DAY_OF_MONTH                                     18         12          7
OP_CARRIER                                       WN         WN         WN
ORIGIN                                          DEN        DEN        DEN
ORIGIN_STATE_NM                             Florida California    Florida
DEST                                            DEN        DEN        ATL
DEST_STATE_NM                               Florida    Florida    Florida
DEP_DELAY_GROUP                                   2          3          1
ARR_DELAY_GROUP                                   1          3          1
DISTANCE_GROUP                                    2          2          2
CARRIER_DELAY                               29.7853    51.0329    17.9748
```

K-Means (n =4)

```
Initial starting points (random):

Cluster 0: 4,12,UA,SFO,California,IAD,Virginia,3,3,10,45,0,2,0,0
Cluster 1: 4,10,QX,SEA,Washington,BLI,Washington,1,1,1,2,0,0,0,21
Cluster 2: 2,18,NK,PHX,Arizona,MSP,Minnesota,3,3,6,0,0,13,0,34
Cluster 3: 1,1,WN,DEN,Colorado,MTJ,Colorado,5,6,1,10,0,16,0,71

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                    Cluster#
Attribute                              Full Data          0          1          2          3
                                      (406372.0)  (62908.0) (157253.0)  (75638.0) (110573.0)
=============================================================================================
MONTH                                         4          4          4          2          1
DAY_OF_MONTH                                 18         12         10         18          1
OP_CARRIER                                   WN         UA         WN         WN         WN
ORIGIN                                      DEN        LAX        ATL        PHX        DEN
ORIGIN_STATE_NM                         Florida California    Florida    Florida   Colorado
DEST                                        DEN        LAX        ATL        ORD        DEN
```

K-Means (n =6)

```
Number of iterations: 28
Within cluster sum of squared errors: 2973443.854241131

Initial starting points (random):

Cluster 0: 4,12,UA,SFO,California,IAD,Virginia,3,3,10,45,0,2,0,0
Cluster 1: 4,10,QX,SEA,Washington,BLI,Washington,1,1,1,2,0,0,0,21
Cluster 2: 2,18,NK,PHX,Arizona,MSP,Minnesota,3,3,6,0,0,13,0,34
Cluster 3: 1,1,WN,DEN,Colorado,MTJ,Colorado,5,6,1,10,0,16,0,71
Cluster 4: 4,26,WN,MIA,Florida,BWI,Maryland,0,4,4,9,0,57,0,3
Cluster 5: 1,9,WN,AUS,Texas,LAS,Nevada,2,1,5,9,0,0,0,17

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                    Cluster#
Attribute                                Full Data          0          1          2          3          4          5
                                        (406372.0)  (52000.0)  (81188.0)  (52730.0)  (46329.0)  (61476.0) (112649.0)
=============================================================================================================================
MONTH                                            4          4          4          2          1          4          1
DAY_OF_MONTH                                    18         12         10         18          1          7          9
OP_CARRIER                                      WN         UA         DL         AA         WN         WN         WN
ORIGIN                                         DEN        LAX        ATL        PHX        DEN        MCO        DFW
ORIGIN_STATE_NM                          Florida,Cal  California  Georgia,  Arizona,  Colorado,  Florida,  Texas,
```

K-Means (n =8)

```
Initial starting points (random):

Cluster 0: 4,12,UA,SFO,California,IAD,Virginia,3,3,10,45,0,2,0,0
Cluster 1: 4,10,QX,SEA,Washington,BLI,Washington,1,1,1,2,0,0,0,21
Cluster 2: 2,18,NK,PHX,Arizona,MSP,Minnesota,3,3,6,0,0,13,0,34
Cluster 3: 1,1,WN,DEN,Colorado,MTJ,Colorado,5,6,1,10,0,16,0,71
Cluster 4: 4,26,WN,MIA,Florida,BWI,Maryland,0,4,4,9,0,57,0,3
Cluster 5: 1,9,WN,AUS,Texas,LAS,Nevada,2,1,5,9,0,0,0,17
Cluster 6: 1,29,QX,ANC,Alaska,FAI,Alaska,0,1,2,8,0,9,0,0
Cluster 7: 3,17,YX,LGA,'New York',PNS,Florida,12,12,5,0,0,28,0,179

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                          Cluster#
Attribute                                      Full Data          0          1          2          3          4          5          6
                                              (406372.0)  (41705.0)  (69061.0)  (48413.0)  (31665.0)  (45039.0)  (70029.0)  (56639.0)  (4
-------------------------------------------------------------------------------------------------------------------------------------------
```

K-Means (n =10)

```
Initial starting points (random):

Cluster 0: 4,12,UA,SFO,California,IAD,Virginia,3,3,10,45,0,2,0,0
Cluster 1: 4,10,QX,SEA,Washington,BLI,Washington,1,1,1,2,0,0,0,21
Cluster 2: 2,18,NK,PHX,Arizona,MSP,Minnesota,3,3,6,0,0,13,0,34
Cluster 3: 1,1,WN,DEN,Colorado,MTJ,Colorado,5,6,1,10,0,16,0,71
Cluster 4: 4,26,WN,MIA,Florida,BWI,Maryland,0,4,4,9,0,57,0,3
Cluster 5: 1,9,WN,AUS,Texas,LAS,Nevada,2,1,5,9,0,0,0,17
Cluster 6: 1,29,QX,ANC,Alaska,FAI,Alaska,0,1,2,8,0,9,0,0
Cluster 7: 3,17,YX,LGA,'New York',PNS,Florida,12,12,5,0,0,28,0,179
Cluster 8: 1,2,YX,ORF,Virginia,ORD,Illinois,2,2,3,0,0,3,0,32
Cluster 9: 2,3,NK,BOS,Massachusetts,FLL,Florida,3,3,5,0,0,2,0,53

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                        Cluster#
Attribute                                  Full Data            0            1            2            3            4
                                          (406372.0)   (33670.0)    (61428.0)    (30796.0)    (26545.0)    (42462.0)
==========================================================================================================================
```

Al crear clústeres con este algoritmo, se pudo observar que la cantidad adecuada para la formación de clústeres con el algoritmo K-Means es de 10.

En cuanto al modelo generado, los resultados concernientes son:

- Total de iteraciones: 3
- Puntos iniciales (Aleatorios):

Cluster 0: 4,12,UA,SFO,California,IAD,Virginia,3,3,10,45,0,2,0,0

Cluster 1: 4,10,QX,SEA,Washington,BLI,Washington,1,1,1,2,0,0,0,21

Cluster 2: 2,18,NK,PHX,Arizona,MSP,Minnesota,3,3,6,0,0,13,0,34

Cluster 3: 1,1,WN,DEN,Colorado,MTJ,Colorado,5,6,1,10,0,16,0,71

Cluster 4: 4,26,WN,MIA,Florida,BWI,Maryland,0,4,4,9,0,57,0,3

Cluster 5: 1,9,WN,AUS,Texas,LAS,Nevada,2,1,5,9,0,0,0,17

Cluster 6: 1,29,QX,ANC,Alaska,FAI,Alaska,0,1,2,8,0,9,0,0

Cluster 7: 3,17,YX,LGA,'New York',PNS,Florida,12,12,5,0,0,28,0,179

Cluster 8: 1,2,YX,ORF,Virginia,ORD,Illinois,2,2,3,0,0,3,0,32

Cluster 9: 2,3,NK,BOS,Massachusetts,FLL,Florida,3,3,5,0,0,2,0,53

- Instancias de los clústeres:

0 33670 (8%)

1 61428 (15%)

2 30796 (8%)

3 26545 (7%)

4 42462 (10%)

5 56090 (14%)

6 49577 (12%)

7 38573 (9%)

8 39723 (10%)

9 27508 (7%)

- Centroides finales:

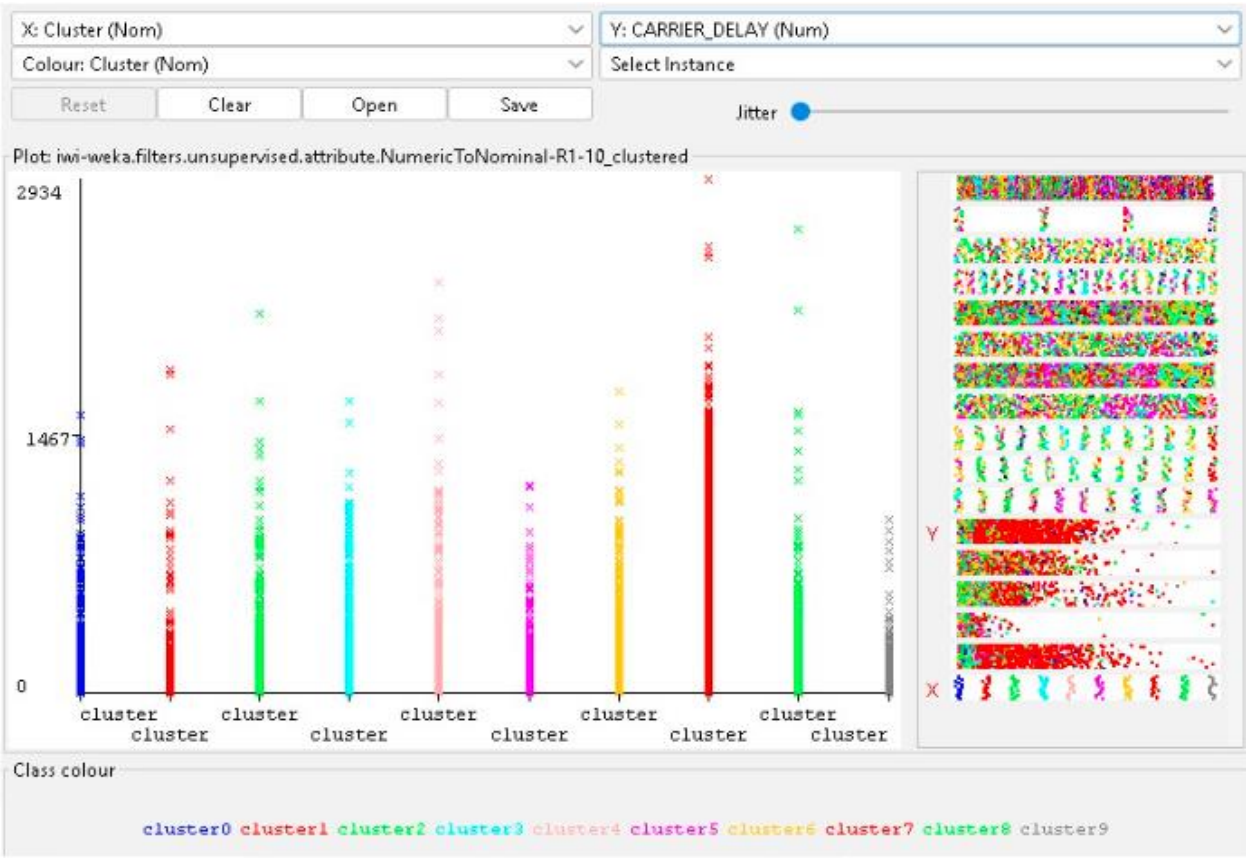
Final cluster centroids:

Attribute	Full Data (406372.0)	Cluster# 0 (33670.0)	1 (61428.0)	2 (30796.0)	3 (26545.0)	4 (42462.0)	5 (56090.0)	6 (49577.0)	7 (38573.0)	8 (39723.0)
MONTH	4	4	4	2	1	4	1	1	3	1
DAY_OF_MONTH	18	12	10	18	1	13	9	7	17	2
OP_CARRIER	WN	UA	WN	WN	WN	WN	WN	OO	B6	YX
ORIGIN	DEN	SFO	SEA	PHX	DEN	MCO	DFW	ORD	LGA	DCA
ORIGIN_STATE_NM	Florida	California	Washington	Arizona	Colorado	Florida	Texas	California	New York	Virginia M
DEST	DEN	LAX	SEA	MSP	DEN	DFW	LAS	ATL	MCO	ORD
DEST_STATE_NM	Florida	California	California	Minnesota	Colorado	Texas	Texas	California	Florida	Illinois
DEP_DELAY_GROUP	2	3	1	3	5	0	2	0	12	2
ARR_DELAY_GROUP	1	3	1	3	6	4	1	1	12	2
DISTANCE_GROUP	2	2	1	3	1	4	5	2	5	3
CARRIER_DELAY	29.7853	18.3575	16.5169	39.5306	29.3669	22.3934	17.8149	25.3253	97.1979	20.7919
WEATHER_DELAY	3.9667	1.6653	1.6847	3.7274	2.8275	3.4913	2.166	7.0099	10.4755	3.8185
NAS_DELAY	10.1968	5.7812	4.5609	5.3751	7.2284	19.0454	3.934	11.0431	19.7147	10.7539
SECURITY_DELAY	0.1579	0.0998	0.1954	0.1708	0.1044	0.1684	0.1782	0.1038	0.2006	0.1493
LATE_AIRCRAFT_DELAY	27.5473	40.3012	8.4238	20.0922	49.8788	32.4067	15.5415	8.3996	83.7245	25.5891

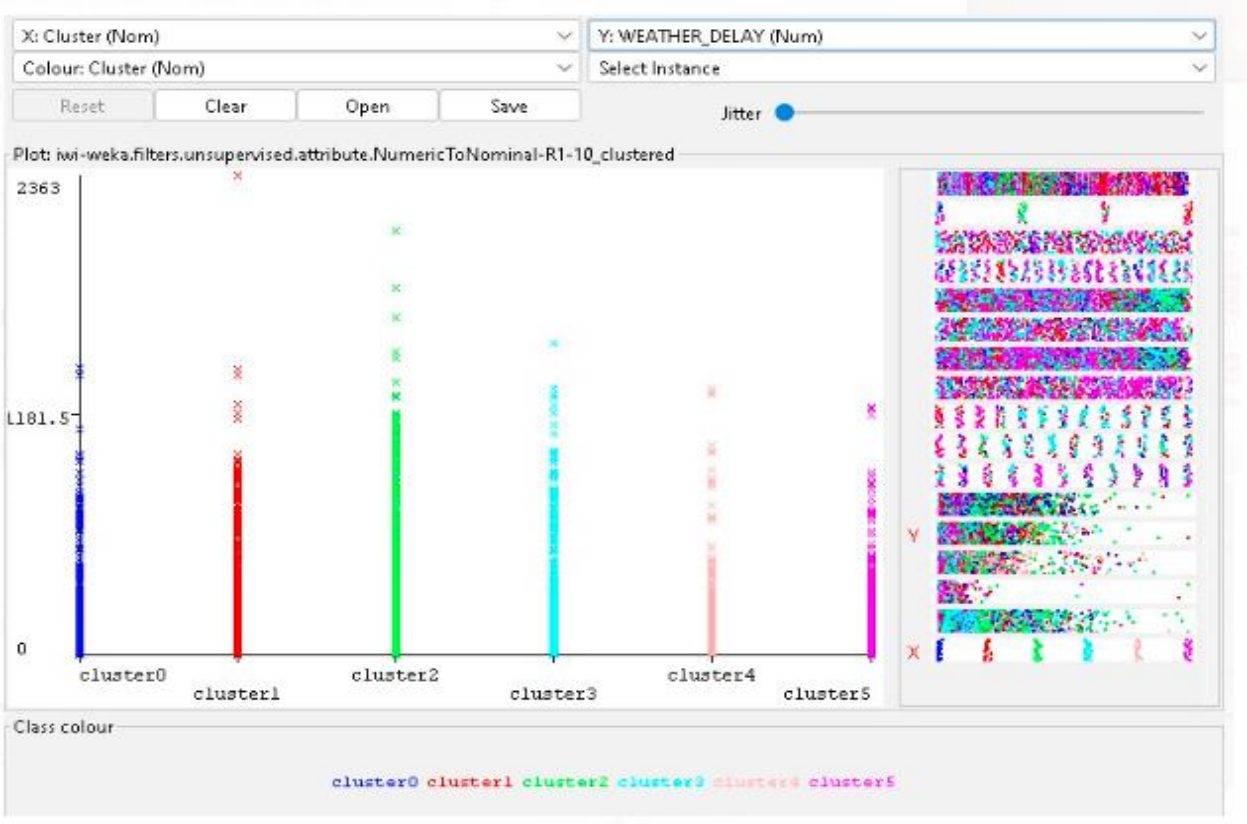
- Mostrar distribución de clústeres (Gráfico):

Aquí se tomaron en cuenta los atributos concernientes a los retrasos en los vuelos, los cuales serán pilares para el cumplimiento de los objetivos propuestos.

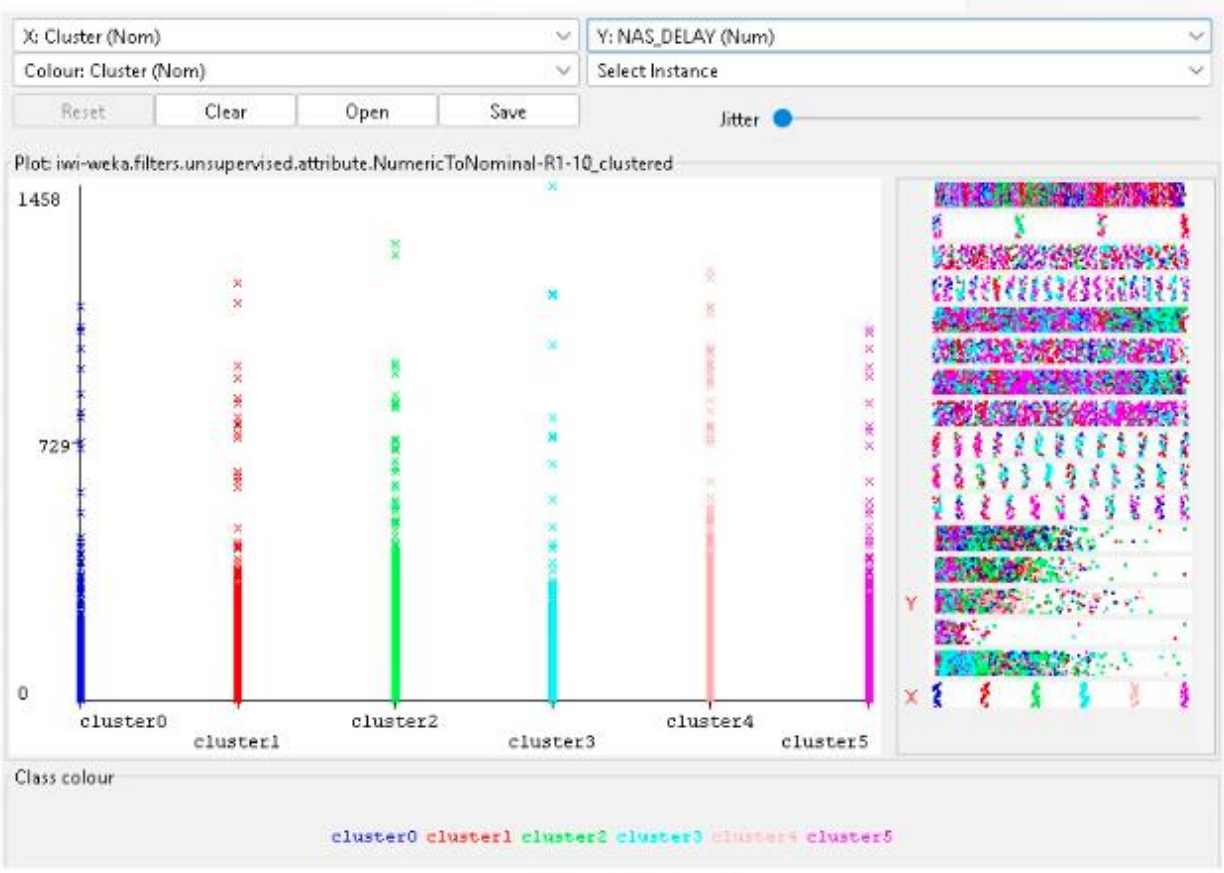
Cluster vs Carrier Delay



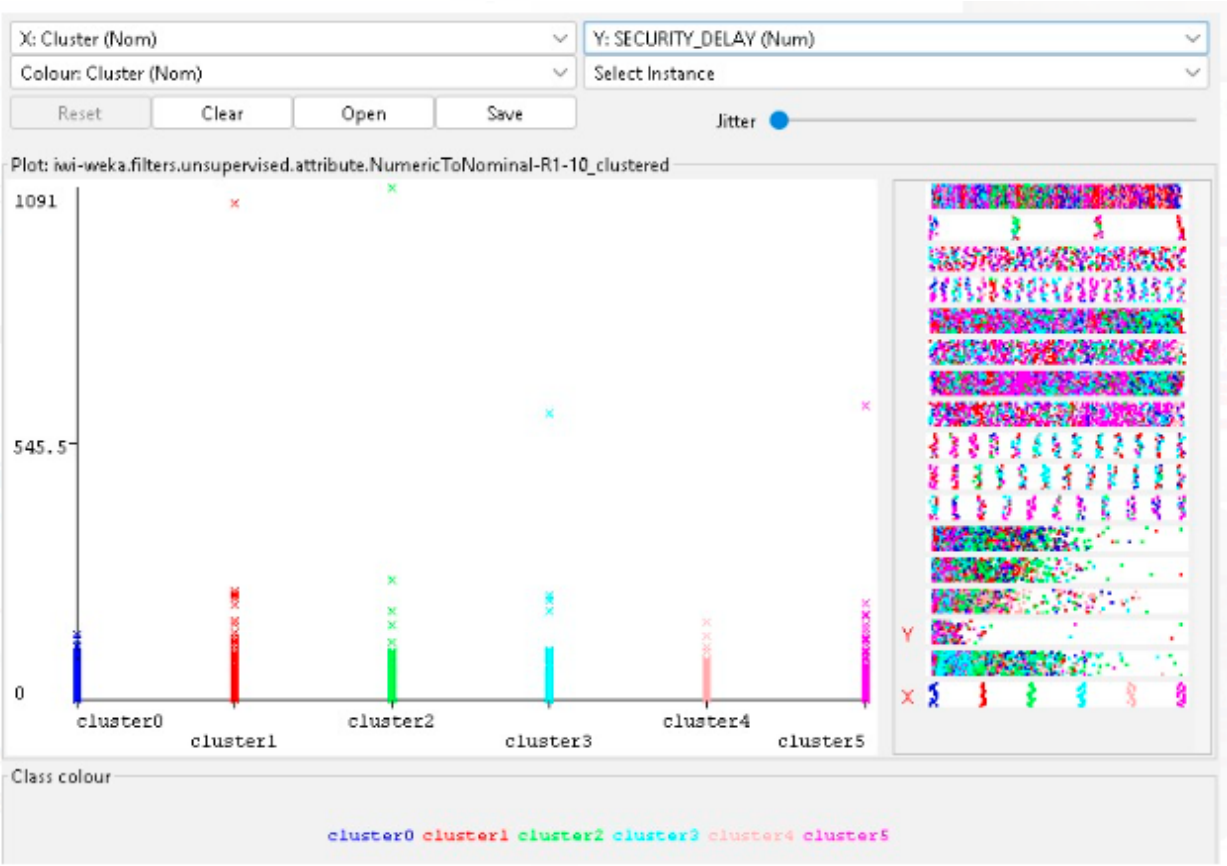
Cluster vs Weather Delay



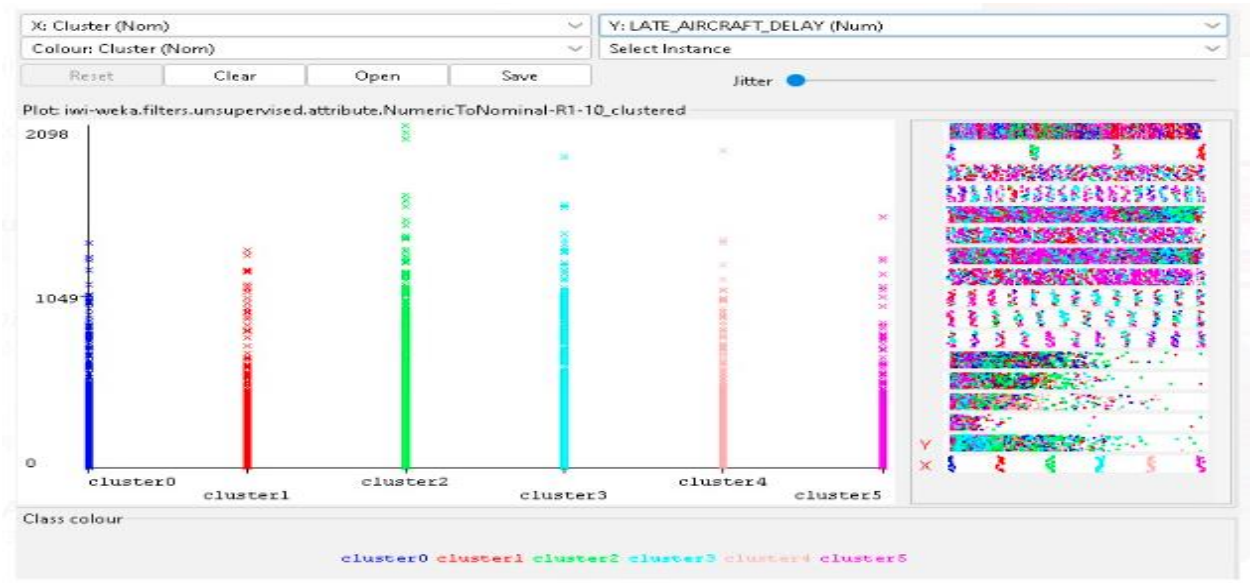
Cluster vs Nas Delay



Cluster vs Security Delay



Cluster vs Aircraft Delay



En cuanto a las características del modelo, podemos decir que en todos los clústeres se presenta algo en común: los retrasos debido a carrier son de los más altos en consideración a los otros tipos de retraso. Por otro lado, los retrasos de seguridad son siempre muy bajos, comparándolos con los demás. En particular, el clúster 7 es el que tiene más minutos acumulados en todos los tipos de retrasos de los vuelos, a pesar de no ser el de mayor cantidad de instancias, y es el único que abarca el mes 3, es decir marzo.

4.4 Evaluar el Modelo.

En este modelo, se encontraron resultados que revelan clústeres donde se agrupan las variedades según la similitud de sus atributos numéricos. Estos atributos se refieren a CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY y LATE_AIRCRAFT_DELAY.

Con el modelo de K-Means fue posible encontrar 10 grupos de variedades en las que la principal diferencia entre las 10 es respecto a la ocurrencia y duración de los retrasos.

Evaluación

5.1 Evaluar los resultados

Evaluando los resultados de la minería de datos con respecto a los criterios de éxito empresarial.

Con correlación a los resultados de la fase de modelado, usando la técnica de clustering para reunir en grupos con variación en sus retrasos, de la cual se puede notar las diferencias que hay entro los atributos de los tipos de retrasos.

Teniendo un enfoque al objetivo inicial de minería de datos propuesto en la fase de entendimiento del negocio, se pretende obtener un conjunto de clústeres que ayuden a determinar las características comunes que influyen en los retrasos de vuelos aéreos, y de igual forma otros datos complementarios como las características específicas de los vuelos, y así predecir con mayor facilidad estas eventualidades. Al lograr conocer sobre cuáles de estos retrasos tienen mayor ocurrencia y cuestan mayor cantidad de minutos, es posible temar medidas sobre los aeropuertos que puedan mejorar la calidad del servicio brindada a los clientes.

Cabe destacar que, con los modelos obtenidos utilizando el algoritmo K-Means se cumplen todos los criterios de éxito ya establecidos con el objetivo antes mencionado. Cumpliéndose de la siguiente manera:

Se logró agrupar al menos 10 clústeres bien definidos y observar la relación que poseen los elementos. D esta manera, la cercanía que tienen los elementos de cada clúster hechos por el algoritmo cuando el número de clúster es 10, se nota con más detalles en los atributos que tienen que ver con los tipos de retrasos. Lo cual fue punto clave para que los resultados dieran éxito a lo que el objetivo planteaba.

Se logró determinar un clúster de un mes en donde presenta la mayor cantidad de ocurrencias y costo de tiempo en donde pasan los distintos tipos de retrasos, a pesar de no ser el que más instancias poseía, pero siendo el poseedor de una mejor agrupación de datos, ya que sus atributos eran los que más correlación tenían entre sí.

5.2 Revisión el proceso

Vistazo general del proceso de minería de datos.

Se llevó a cabo un seguimiento preciso de la metodología CRISP-DM para proyectos de minería de datos y se abarcaron todos sus acápites: entendimiento de negocio, entendimiento de la data, preparación de data, modelado, evaluación y deployment.

Análisis del proceso de minería.

Cada etapa de la metodología fue necesaria para llevar a cabo el proyecto mediante una metodología ordenada, probando la utilidad de este estándar llamado CRISP-DM. A la hora de tratar de implementar el algoritmo K-Means se utilizó finalmente la herramienta Weka, puesto que aunque se intentó implementar otras herramientas como R Studio, este no puede manejar la cantidad de datos que se pretendía incluir en el dataset, y debido a la restricción de su memoria se eliminó como herramienta.

Maneras de mejorar.

Debido a lo nuevo de realizar un proyecto de este tipo, la falta de experiencia se pudo en ocasiones apreciar, y a veces se interpretaba incorrectamente que era lo que se pedía al momento de seguir la metodología CRISP-DM. Estas incertidumbres fueron aclaradas mediante fuentes online complementarias, y consultas a compañeros. La manera de solucionar esto es estudiando nuevamente y cuidadosamente la metodología y practicando, llevando a cabo más proyectos

utilizando la metodología, aprendiendo de errores y volviendo sobre los pasos realizados cuantas veces sea necesario.

5.3 Determinar siguientes pasos

Luego de observar los resultados obtenidos por nuestros modelos podemos concluir que estos han sido realmente satisfactorios para los propósitos planteados, y de igual forma para la realización de esta tarea de minería de datos. Además, podemos afirmar que las pruebas realizadas a este modelo resultaron ser adecuadas para el propósito final, tomando todo lo anterior en cuenta se llega a la decisión que no es necesario una o varias implementaciones más para lograr que se cumpla lo planteado a lo largo del proyecto. Sin embargo, si se toman en cuenta otros objetivos, esto daría lugar a nuevos modelos, nuevas transformaciones de datos y en definitiva, más análisis e informaciones, por lo que se podrían evaluar otros propósitos.

Por otro lado, restaría producir y entregar un reporte al Departamento de Transportes de los Estados Unidos detallando los resultados que incluyan nuestro análisis y descubrimientos, como también entregar la data utilizada con fines de que se tenga la libertad de realizar su propio análisis, o simplemente para que se tenga acceso a la fuente que originó todas las conclusiones mencionadas.

Finalmente, se puede decir que la decisión más conveniente es presentar el proyecto y sus conclusiones como tal, porque son suficientes para que se tomen medidas que ayuden en la calidad del servicio aéreo.

Deployment

6.1 Plan de Despliegue

Se planifica la entrega de un documento detallado sobre la recomendación de dar prioridad según por ocurrencias los tipos de retrasos que puedan suceder en un mes, a partir de los hallazgos en los resultados de los modelos realizados, para ser recibido por el departamento de transporte.

Para la documentación, será realizada mediante el análisis de los datos de los vuelos que han ocurrido hasta el mes más actual que brindó la página del departamento de transporte, con los modelos realizados a partir de estos con sus transformaciones. Una vez ya analizado, se va a detallar sobre el tipo de retraso y el costo de tiempo que tardan en despegar.

Además, para cada retraso en el documento se incluirá los datos que conllevan para la razón de su selección apoyándose de los modelos relevantes.

Para finalizar, es contentar con la subsecretaria de investigación y tecnología sobre lo encontrado y el envío de la documentación, esperando a su respuesta y reacción del mismo y compartiéndolo con todo el personal del departamento de transporte y de igual forma a las aerolíneas correspondientes.

6.2 Plan de Monitoreo y Mantenimiento

Para el monitoreo, se planifica observar el efecto que podría causar a seguir las recomendaciones del detallado sobre el costo de tiempo de cada retraso, el cual se hará mediante encuestas no solo al cliente sino al personal que lo sigan. De estas encuestas observar si hubo alguna reducción de tiempo sobre los retrasos que presentan mayor ocurrencia y malgasto del tiempo.

Por otro lado, será recopilar nuevos datos por mes si hay algún nuevo tipo de retraso o incluso, con las expectativas altas, la eliminación de uno, ya que con el avance tecnológico y cambios humanos podrían resultar en variaciones en los atributos obtenidos y por ello conllevaría a rehacer los modelos.

6.3 Realizar Reporte Final

Situación

Una de las secretarías del DOT se ha comunicado con nosotros, dos ingenieros especializados en crear modelos de predicción y análisis de datos dentro de lo que es la minería de datos, expresando el deseo del Departamento de Transportes de los Estados Unidos de poder tener mejor control de la situación de retrasos en los vuelos.

Análisis de negocio

• Antecedentes: El Departamento de Transportes de los Estados Unidos lleva un control detallado de los vuelos retrasados que se realizan dentro de su territorio. Estas estadísticas son recogidas en distintos reportes a lo largo del tiempo, desde el 2011 hasta la actualidad.

• Objetivo de negocio: Mejorar la satisfacción del cliente aéreo y la calidad del servicio de las líneas aéreas.

• Criterio de éxito: Recopilar estadísticas que son recogidas en distintos reportes detallando los vuelos retrasados que se realizan dentro de su territorio, a fin de obtener informaciones relevantes para la mejora de los servicios.

Público objetivo

• Departamento de Transportes de los Estados Unidos

• Aereopuertos

• Aereolíneas

Minería de datos

Objetivos:

• Utilización de una técnica de clustering con la finalidad de agrupar elementos y clasificarlos en conjuntos a partir de sus cualidades, para así determinar las causas comunes que influyen en los retrasos de vuelos aéreos, y de igual forma otros datos complementarios como las características específicas de los vuelos, y así predecir con mayor facilidad estas eventualidades.

Proceso:

Se prepararon los datos para realizar la minería de datos mediante el estudio del dataset obtenido del Bureau de Estadística, con lo cual se pudo procesar y transformar los datos a entradas aceptables y adecuadas para los métodos de minería seleccionados, empleando 15 de 119 de los atributos originales, y 406372 de las 2254521 instancias.

El método de minería seleccionado se basa en la formación de modelos en donde se agrupan los vuelos en conjuntos, se decidió utilizar clustering y analizar el algoritmo K-means mediante la herramienta Weka.

Resultados:

Con los modelos formados, se pudo cumplir el objetivo de minería, debido a que se formaron varios conjuntos y se determinaron que K-means y COBWEB producen conjuntos de elementos con la cual se puede seleccionar uno como lista de recomendación aceptables, esto es debido a que, en ambos, producen un conjunto con el radio más alto de calorías promedio del conjunto por un precio razonable.

Conclusión:

Con los objetivos de negocio y de minería asignados, mediante la ejecución del proyecto se ha logrado cumplir ambas, ya que con el objetivo de minería se pudo determinar un conjunto de características bajo las cuales se obtiene una propensión a incurrir en alguno de los retrasos planteados en los atributos utilizados, por lo que se pueden tomar medidas para que estos retrasos puedan ser previstos con mayor antelación, y dar lugar a planes de repuesto más rápidos por parte de los aeropuertos y aerolíneas, y no sólo prever, sino incluso evitar retrasos, mejorando así la satisfacción del cliente, cumpliendo de tal manera el objetivo de negocio.

6.4 Review Project

Durante el desarrollo del proyecto, el equipo de trabajo utilizó la guía del libro CRISP-DM, el cual tiene una metodología que define un esquema riguroso para un proyecto de minería de datos. El objetivo de este trabajo, fue el de aplicar todo lo aprendido en la asignatura de “Minería de Datos” de la Pontificia Universidad Católica Madre Y Maestra (PUCMM).

Con el dataset proporcionado pudimos trabajar de forma eficiente con el método de minería de datos Kmeans, claramente habiendo preparado los datos anteriormente para estos fines, además haciendo uso de la herramienta Weka, la cual es conocida ampliamente por su funcionalidad con respecto a la minería de datos.

Gracias a esto, se pudo conocer cuáles son los pasos que deben seguirse al momento de llevar a cabo un proyecto de minería de datos en un entorno más práctico y aplicado al mundo actual. Finalmente, resultó evidenciable el notable uso e impacto que tiene la minería de datos en la toma de decisiones que pueden significativamente mejorar la satisfacción, calidad, y el acceso a servicios y productos.