

Análisis de vuelos de 2015 de EU utilizando el sistema de base de datos MONGO DB



**Análisis de datos
Introducción a base de datos
William Hernández Rosas**

Objetivo

- Conocer y ubicar la situación de las aerolíneas estadounidenses.
- Comparar dos fuentes de información de vuelos de EU de 2015. Por un lado, datos de retrasos y cancelación de vuelos de 2015; por otro lado, datos de Twitter sobre análisis de sentimientos de viajeros de EU durante febrero de 2015.

Contexto

- Para este proyecto se obtuvo una muestra de tamaño 500,000 de los datos de retrasos y cancelaciones de vuelos de 2015 y datos de análisis de sentimientos en Twitter de las aerolíneas estadounidenses recopilados del 17 al 24 de febrero 2015 provenientes de la plataforma Kaggle a través de la página <https://www.kaggle.com/>

Collection Name ^	Documents	Avg. Document Size	Total Document Size	Num. Indexes	Total Index Size
Twitter_aerolineas	14,640	538.0 B	7.5 MB	1	156.0 KB
airlines	14	74.4 B	1.0 KB	1	20.0 KB
airports	322	175.8 B	55.3 KB	1	20.0 KB
catalogo_dia_semana	7	53.6 B	375.0 B	1	20.0 KB
sample_flights	500,000	235.5 B	112.3 MB	1	4.6 MB

Planteamiento del problema

¿En qué aerolíneas se debe volar para evitar retrasos significativos?

Solución 1. Obtener la frecuencia de vuelos con retraso en su llegada por compañía

☰

\$project

🔴

🗑️

+

```
1 {
2   AIRLINE: 1,
3   retraso_salida: {
4     $cond: {
5       'if': {
6         $gt: [
7           'DEPARTURE_DELAY',
8           0
9         ]
10      },
11      then: 1,
12      'else': 0
13    }
14  },
15  retraso_llegada: {
16    $cond: {
17      'if': {
18        $gt: [
19          'ARRIVAL_DELAY',
20          0
21        ]
22      },
23      then: 1,
24      'else': 0
25    }
26  }
27 }
```

Output after [\\$project](#) stage ⓘ (Sample of 20 documents)

```
_id: ObjectId("5f26def4bd7cf51340488e00")
AIRLINE: "US"
retraso_salida: 0
retraso_llegada: 0
```

```
_id: ObjectId("5f26def4bd7cf51340488e01")
AIRLINE: "EV"
retraso_salida: 0
retraso_llegada: 0
```

Solución 1. Obtener la frecuencia de vuelos con retraso en su llegada por compañía

|||

▼

\$group

☒

+

Output after [\\$group](#) stage ⓘ (Sample of 14 documents)

```
1 {
2   _id: '$AIRLINE',
3   Total_retraso_Salida: {
4     $sum: '$retraso_salida'
5   },
6   Total_retraso_Llegada: {
7     $sum: '$retraso_llegada'
8   }
9 }
```

_id: "EV"

Total_retraso_Salida: 2919

Total_retraso_Llegada: 3650

_id: "UA"

Total_retraso_Salida: 4399

Total_retraso_Llegada: 3161

|||

▼

\$lookup

☒

+

Output after [\\$lookup](#) stage ⓘ (Sample of 14 documents)

```
1 {
2   from: 'airlines',
3   localField: '_id',
4   foreignField: 'IATA_CODE',
5   as: 'airlines'
6 }
```

_id: "AS"

Total_retraso_Salida: 757

Total_retraso_Llegada: 952

▶ airlines: Array

_id: "MQ"

Total_retraso_Salida: 1543

Total_retraso_Llegada: 1705

▶ airlines: Array

Solución 1. Obtener la frecuencia de vuelos con retraso en su llegada por compañía

The image shows a data transformation pipeline with two stages, both using the `$addFields` stage.

Stage 1:

- Configuration:** `$addFields` stage, toggle is on.
- Input (Sample of 14 documents):**

```
1 {
2   airlines_obj: {
3     $arrayElemAt: [
4       '$airlines',
5       0
6     ]
7   }
8 }
```
- Output (Sample of 14 documents):**
 - Document 1: `_id: "F9", Total_retraso_Salida: 597, Total_retraso_Llegada: 711`
 - Document 2: `_id: "NK", Total_retraso_Salida: 873, Total_retraso_Llegada: 931`

Stage 2:

- Configuration:** `$addFields` stage, toggle is on.
- Input (Sample of 14 documents):**

```
1 {
2   airline: '$airlines_obj.AIRLINE'
3 }
```
- Output (Sample of 14 documents):**
 - Document 1: `_id: "VX", Total_retraso_Salida: 426, Total_retraso_Llegada: 448`
 - Document 2: `_id: "NK", Total_retraso_Salida: 873, Total_retraso_Llegada: 931`

Solución 1. Obtener la frecuencia de vuelos con retraso en su llegada por compañía

The screenshot displays the MongoDB Atlas aggregation pipeline editor. It shows two stages: **\$project** and **\$sort**. The **\$project** stage is active, and its output is shown in the right-hand pane. The **\$sort** stage is also active, and its output is shown in the right-hand pane. The left-hand pane shows the JSON documents being processed at each stage.

Stage 1: \$project

Output after **\$project** stage (Sample of 14 documents)

```
1 {
2   _id: 0,
3   airline: 1,
4   Total_retraso_Llegada: 1
5 }
```

Two sample documents are shown in the output pane:

```
Total_retraso_Llegada: 1693
airline: "JetBlue Airways"
```

```
Total_retraso_Llegada: 531
airline: "Hawaiian Airlines Inc."
```

Stage 2: \$sort

Output after **\$sort** stage (Sample of 14 documents)

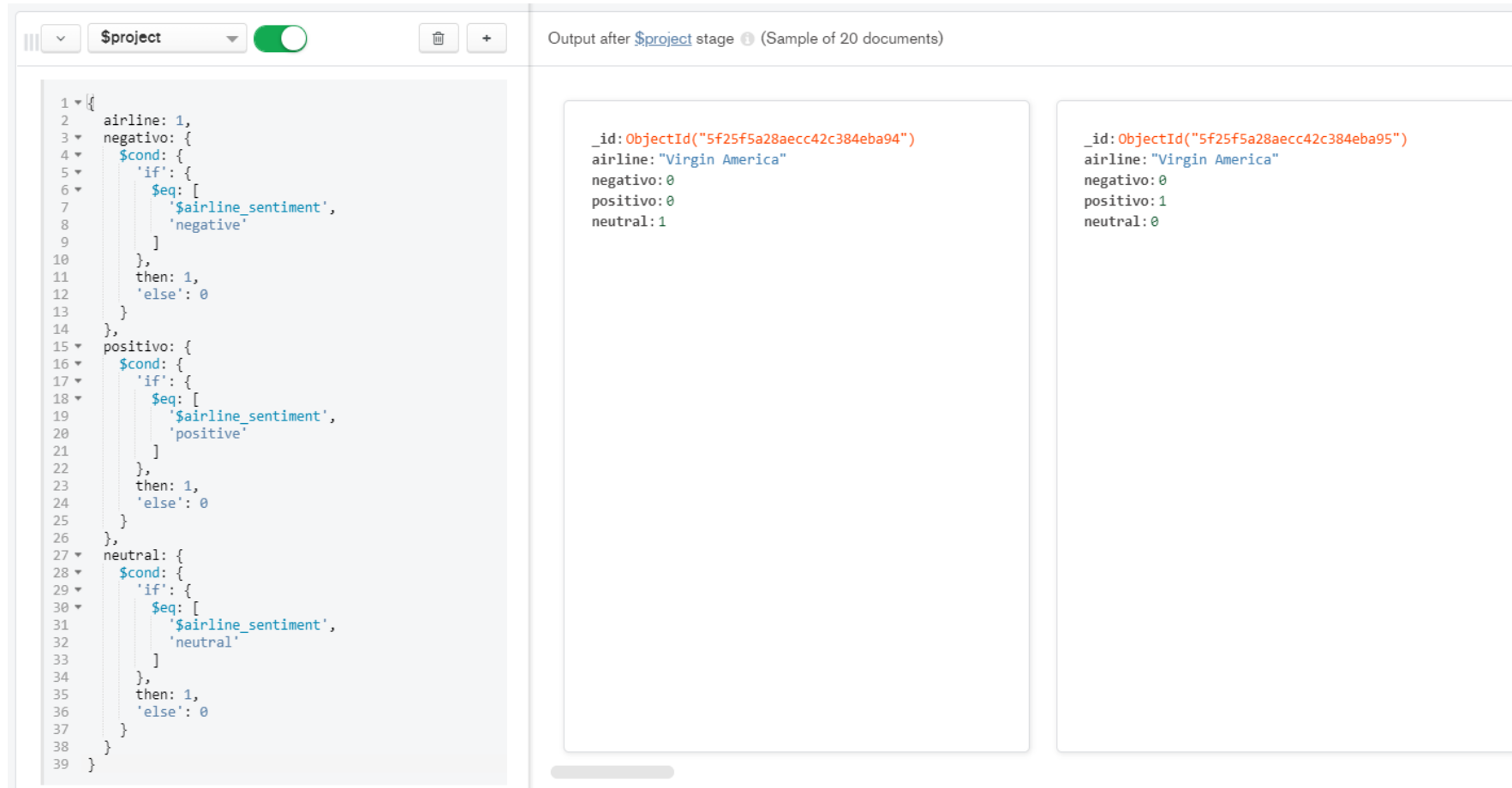
```
1 {
2   Total_retraso_Llegada: -1
3 }
```

Two sample documents are shown in the output pane:

```
Total_retraso_Llegada: 8014
airline: "Southwest Airlines Co."
```

```
Total_retraso_Llegada: 4419
airline: "American Airlines Inc."
```


Solución 2. Creación de un índice de emociones negativas, con apoyo de los datos de análisis de sentimientos en Twitter de las aerolíneas estadounidenses.



The screenshot displays the MongoDB Atlas interface. On the left, a pipeline stage is defined in a code editor. The stage is a \$project operation that creates three new fields: 'negativo', 'positivo', and 'neutral'. Each field is determined by a \$cond operation that checks the value of '\$airline_sentiment' against a list of possible values ('negative', 'positive', 'neutral') and assigns a corresponding value (1 for the match, 0 otherwise).

```
1 {
2   airline: 1,
3   negativo: {
4     $cond: {
5       'if': {
6         $eq: [
7           '$airline_sentiment',
8           'negative'
9         ]
10      },
11      then: 1,
12      'else': 0
13    }
14  },
15  positivo: {
16    $cond: {
17      'if': {
18        $eq: [
19          '$airline_sentiment',
20          'positive'
21        ]
22      },
23      then: 1,
24      'else': 0
25    }
26  },
27  neutral: {
28    $cond: {
29      'if': {
30        $eq: [
31          '$airline_sentiment',
32          'neutral'
33        ]
34      },
35      then: 1,
36      'else': 0
37    }
38  }
39 }
```

On the right, the output of the pipeline stage is shown for a sample of 20 documents. Two documents are visible, both for the airline 'Virgin America'. The first document has 'negativo: 0', 'positivo: 0', and 'neutral: 1'. The second document has 'negativo: 0', 'positivo: 1', and 'neutral: 0'.

```
_id: ObjectId("5f25f5a28aecc42c384eba94")
airline: "Virgin America"
negativo: 0
positivo: 0
neutral: 1

_id: ObjectId("5f25f5a28aecc42c384eba95")
airline: "Virgin America"
negativo: 0
positivo: 1
neutral: 0
```

Solución 2. Creación de un índice de emociones negativas, con apoyo de los datos de análisis de sentimientos en Twitter de las aerolíneas estadounidenses.

The screenshot displays a MongoDB Atlas pipeline editor with two stages. The first stage, `$addFields`, is active and shows a sample of 20 documents. The second stage, `$group`, is also active and shows a sample of 6 documents.

Stage 1: \$addFields

Output after `$addFields` stage (Sample of 20 documents)

```
1 {
2   total: {
3     $sum: 1
4   }
5 }
```

Two sample documents are shown:

```
{
  _id: ObjectId("5f25f5a28aecc42c384eba94"),
  airline: "Virgin America",
  negativo: 0,
  positivo: 0,
  neutral: 1,
  total: 1
}
```

```
{
  _id: ObjectId("5f25f5a28aecc42c384eba95"),
  airline: "Virgin America",
  negativo: 0,
  positivo: 1,
  neutral: 0,
  total: 1
}
```

Stage 2: \$group

Output after `$group` stage (Sample of 6 documents)

```
1 {
2   _id: '$airline',
3   negativos: {
4     $sum: '$negativo'
5   },
6   positivos: {
7     $sum: '$positivo'
8   },
9   neutral: {
10    $sum: '$neutral'
11  },
12  total: {
13    $sum: '$total'
14  }
15 }
```

Two sample documents are shown:

```
{
  _id: "Delta",
  negativos: 955,
  positivos: 544,
  neutral: 723,
  total: 2222
}
```

```
{
  _id: "US Airways",
  negativos: 2263,
  positivos: 269,
  neutral: 381,
  total: 2913
}
```

Solución 2. Creación de un índice de emociones negativas, con apoyo de los datos de análisis de sentimientos en Twitter de las aerolíneas estadounidenses.

The screenshot displays a MongoDB Atlas pipeline editor with two stages. The first stage, **\$addFields**, is active and shows a sample of 6 documents. The second stage, **\$project**, is also active and shows a sample of 6 documents.

Stage 1: \$addFields

Output after **\$addFields** stage (Sample of 6 documents):

```
1 {
2   indice_negativos: {
3     $divide: [
4       '$negativos',
5       '$total'
6     ]
7   }
8 }
```

Sample documents:

- Virgin America**: negativos: 181, positivos: 152, neutral: 171, total: 504, indice_negativos: 0.35912698412698413
- Delta**: negativos: 955, positivos: 544, neutral: 723, total: 2222, indice_negativos: 0.4297929792979298

Stage 2: \$project

Output after **\$project** stage (Sample of 6 documents):

```
1 {
2   _id: 0,
3   aerolinea: "$_id",
4   negativos: 1,
5   positivos: 1,
6   neutral: 1,
7   total: 1,
8   indice_negativos: {
9     $round: [
10      '$indice_negativos', 2
11    ]
12  }
13 }
```

Sample documents:

- Delta**: negativos: 955, positivos: 544, neutral: 723, total: 2222, aerolinea: "Delta", indice_negativos: 0.43
- US Airways**: negativos: 2263, positivos: 269, neutral: 381, total: 2913, aerolinea: "US Airways", indice_negativos: 0.78

Solución 2. Creación de un índice de emociones negativas, con apoyo de los datos de análisis de sentimientos en Twitter de las aerolíneas estadounidenses.



Output after `$sort` stage ⓘ (Sample of 6 documents)

```
1 {  
2   indice_negativos:-1  
3 }
```

<code>negativos: 2263</code> <code>positivos: 269</code> <code>neutral: 381</code> <code>total: 2913</code> <code>aerolinea: "US Airways"</code> <code>indice_negativos: 0.78</code>	<code>negativos: 1960</code> <code>positivos: 336</code> <code>neutral: 463</code> <code>total: 2759</code> <code>aerolinea: "American"</code> <code>indice_negativos: 0.71</code>
---	---

Conclusiones

- La aerolínea con mayor cantidad de retraso en sus llegadas es la aerolínea ***Southwest Airlines Co.*** con 40,548, seguido con 21,800 la aerolínea ***American Airlines Inc.*** y ***Delta Air Lines Inc.*** con 21,505.
- Las principales razones de cancelación de vuelos son por razones de las aerolíneas (errores mecánicos y problemas con los horarios de la tripulación), seguido de las condiciones meteorológicas.
- La mayor cantidad de vuelos se realizan entre los días de semana lunes a viernes.
- El aeropuerto con mayor retraso en sus vuelos tanto en llegadas como salidas es ***Hartsfiels-Jackson Atlanta International.***
- El estado con más aeropuertos es Texas.

Conclusiones

- Las horas de salida programadas abiertas 6:37, 7:23 que no terminan en 0 o 30, son las horas que presentan el mayor promedio de retraso de salida de un avión.
- Las principales razones negativas de viajeros que expresaron sus emociones en Twitter sobre las aerolíneas son: servicio al cliente, vuelo tardío, vuelo cancelado, equipaje perdido, vuelo malo.
- Las Aerolíneas estadounidenses con más tweets negativos son ***United Air Lines Inc.*** y ***US Airways Inc.***
- Se confirma a través de los indicadores propuestos como índice de emociones y razón de proporción negativa que la aerolínea con más atrasos son ***US Airways Inc.*** y ***American Airlines Inc.***
- Finalmente, las aerolíneas menos recomendables para realizar un vuelo por retrasos o cancelaciones son:
 - ***Southwest Airlines Co.***
 - ***American Airlines Inc.***
 - ***Delta Air Lines Inc.***
 - ***US Airways Inc.***