PONTIFICIA UNIVERSIDAD CATÓLICA MADRE Y MAESTRA

Nombre:

Natasha María López Concepción

Vladimir Osvaldo Curiel Ovalles

Matricula:

1014-1274

1014-1415

Docente:

Lisibonny Eustina Beato

Materia:

Inteligencia de Negocios

Titulo:

Informe del Dashboard

Introducción

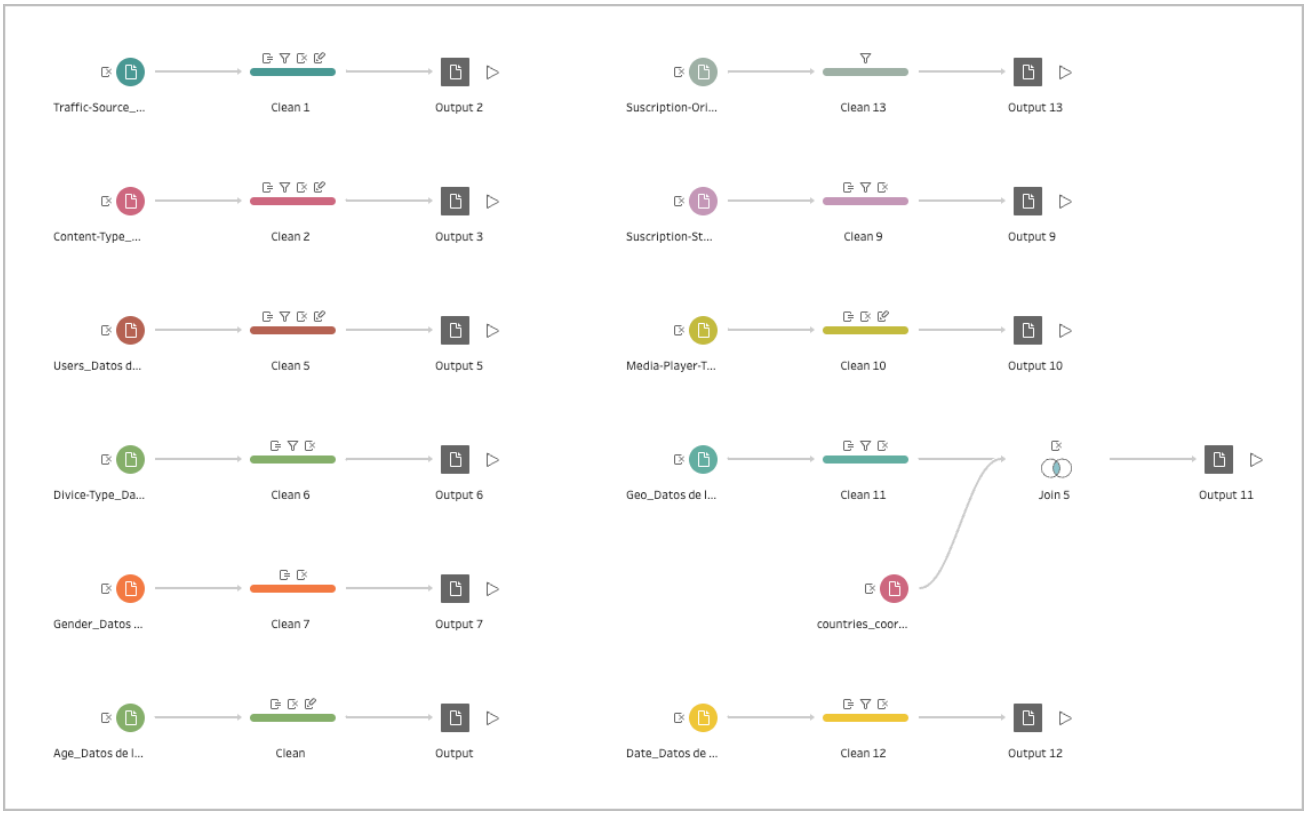
En este informe, se presenta el desarrollo de un dashboard interactivo como parte del proyecto final para la asignatura. El objetivo es analizar las métricas de vistas y suscriptores del canal de YouTube *The Quiz Challenge*, con el fin de identificar los factores que afectan el crecimiento del canal y la interacción con su audiencia. A pesar de su popularidad, el canal enfrenta dificultades para comprender qué contenido genera más vistas y suscriptores, por lo que se busca brindar una herramienta que ayude a tomar decisiones basadas en datos.

El proyecto se enfoca en diseñar e implementar un dashboard que permita visualizar de manera clara las métricas clave, como las vistas y suscriptores, y explorar qué elementos del contenido tienen mayor impacto. Para ello, se llevará a cabo un proceso de preparación y transformación de los datos, seguido de la creación del dashboard interactivo utilizando herramientas de visualización como Tableau. Este informe documentará todo el proceso y presentará los resultados obtenidos.

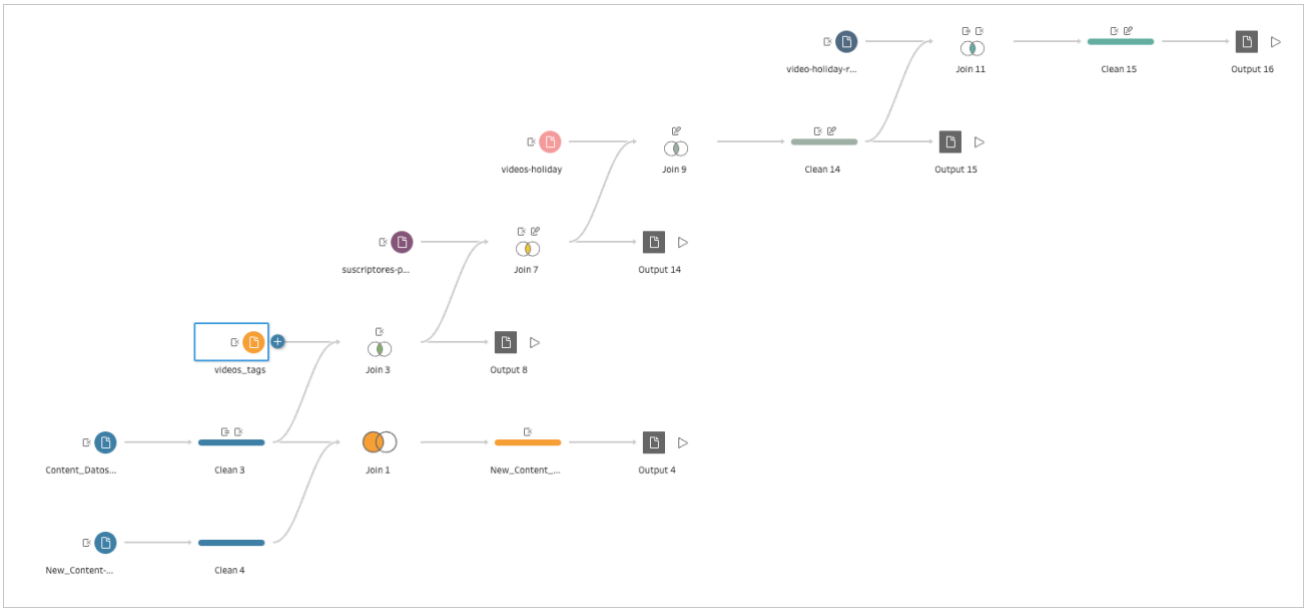
Inicialmente, se nos otorgó una cantidad de información poco relacionada entre sí y sobre todo poco relacionado con los videos **de manera explícita,** por lo que nos embarcamos en la tarea de analizar si verdaderamente la data se puede anexar de alguna forma. El primer paso fue revisar bien a detalles toda la información que contenía el dataset y entender a qué se refiere cada métrica que otorgan las estadísticas de YouTube. Luego la tarea más ardua fue preparar la data y encontrar de qué forma hacer que los datos se puedan vincular con los videos, para ellos, partimos de **asumir** que las métricas obtenidas entre dos fechas se le atribuyen al video más próximo a la fecha inicial. Del mismo modo, para enriquecer la información, fue necesario agregar datos extraídos de la API de YouTube, generar información nueva a partir de la existente utilizando técnicas avanzadas de NLP.

Pre-procesamiento & Transformación de la Información

Lo primero que realizamos fue importar todo la data en ***Tableau Prep*** y realizar una limpieza rápida; eliminando datos que no vamos a usar y cambiando ciertos formatos de datos a unos que podamos sacarle provecho, como estandarizar las duraciones a segundos ya que se encuentran en un formato de fecha.

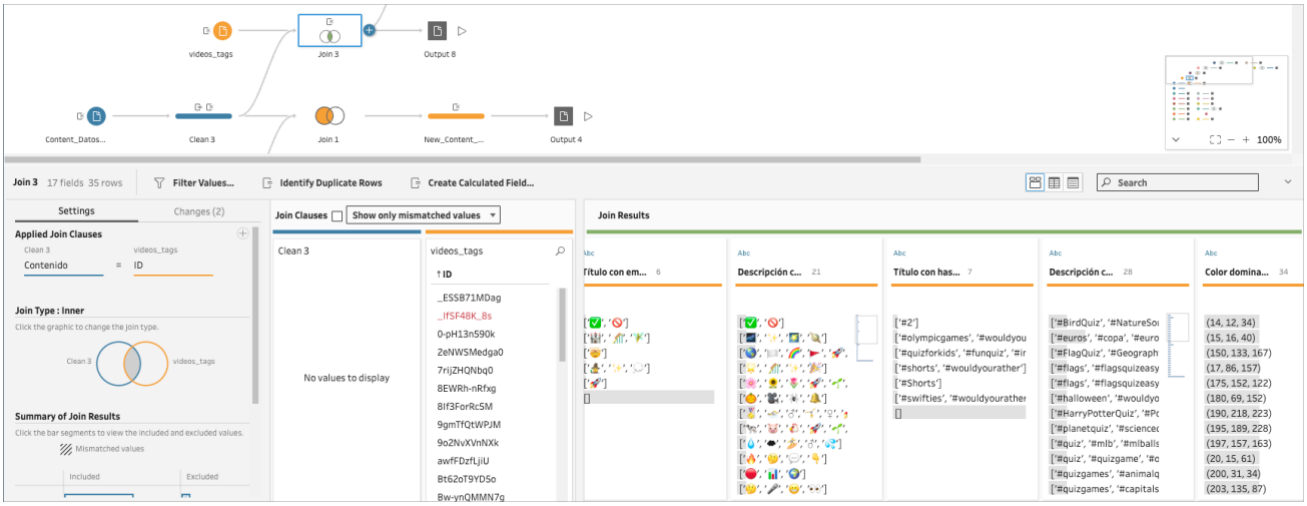


También realizamos una pequeña transformación al CSV de los datos geográficos agregando de manera externas las latitudes y longitudes de cada país presente en el dataset. Cabe destacar que todo eso fue para la información básica que analizamos para intentar sacar una utilidad, pero, la información más importante al final se encontraba en el CSV de los videos, por lo cual es al que más atención le otorgamos y el cual pasó por más transformaciones.



Aquí podemos observar el flujo por el cual pasó el CSV de videos, es importante mencionar que también se realizaron transformaciones de información fuera de Tableau Prep usando Notebooks de Jupyter con Python (Pandas) para agregar data con un carácter más complejo (más que nada para optimizar el tiempo, ya que tareas como las agregaciones y filtros complejos son ciertamente posibles en Tableau Prep, el conocimiento previo nos otorga más soltura en Notebooks usando Pandas)

La primera transformación por la cual pasó el CSV de videos, fue agregarle cosas como: Categoría, Descripción, Duración del Video (Segundos), Emojis en la descripción y el título, Hashtags en la descripción y el titulo y por último el color dominante de las miniaturas de los videos. Eso corresponde a esta parte del flujo:



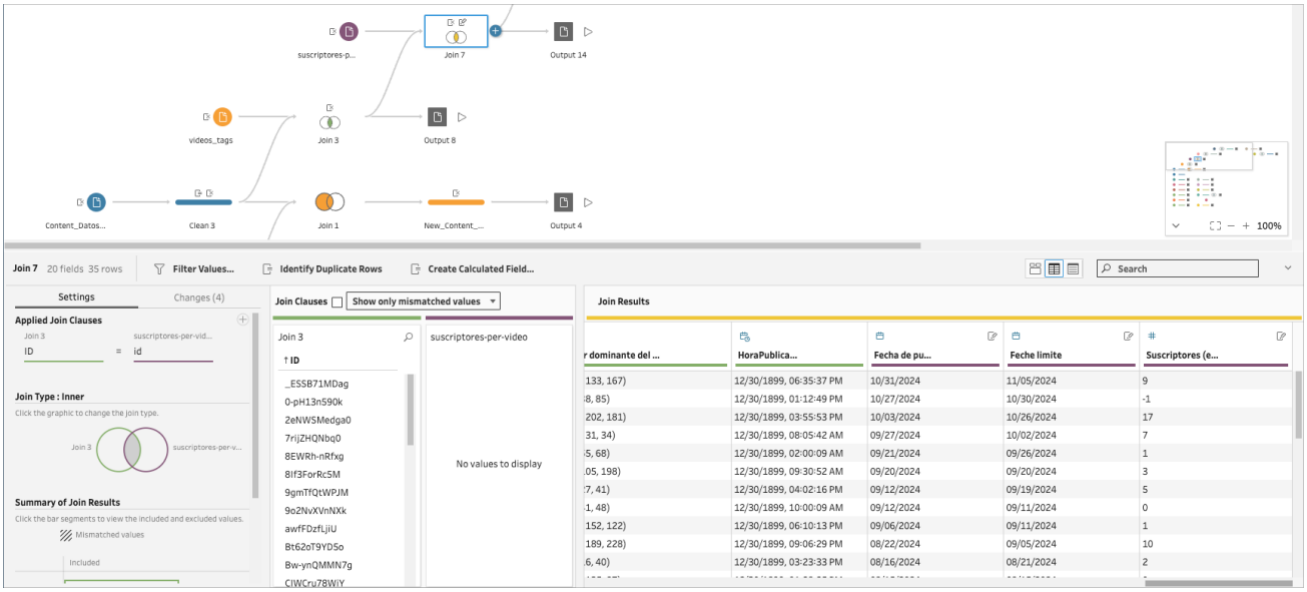
Para obtener esa información extra que teníamos ahí, recurrimos a la API de YouTube y generamos un Notebook

```python
# Procesar datos y crear el CSV
def create_csv():
    videos = get_videos_from_playlist()
    data = []
    for video in videos:
        try:
            video_id = video["snippet"]["resourceId"]["videoId"]
            details = get_video_details(video_id)
            duration = duration_to_seconds(details["contentDetails"]["duration"])
            description = details["snippet"].get("description", "").replace("\n", " ")
            category = get_video_category(details["snippet"]["categoryId"])
            published_at = details["snippet"].get("publishedAt", "")
            published_time = published_at.split("T")[1].replace("Z", "") if "T" in published_at else ""
            thumbnails = details["snippet"]["thumbnails"]
            thumbnail_url = thumbnails.get("maxres", thumbnails.get("standard", thumbnails.get("high", thumbnails.get("medium"))))["url"]
            image = download_image(thumbnail_url)
            dominant_color = get_dominant_color(image)
            titulo_emoji = extract_emojis(details["snippet"]["title"])
            descripción_emoji = extract_emojis(description)
            titulo_hashtags = extract_hashtags(details["snippet"]["title"])
            descripción_hashtags = extract_hashtags(description)
```

Sacando el ID del video, nos podemos asegurar en vincular los videos de manera correcta, las categorías fueron creadas arbitrariamente analizando los videos del canal y las aplicamos con una función en Python

```python
def assign_category(title):
    title = title.lower()
    for category, keywords in categories.items():
        if any(keyword.lower() in title for keyword in keywords):
            return category
    return "Otros"

df['Categoría'] = df['Titulo'].apply(assign_category)

categorized_csv = csv_file.replace(".csv", "_tags.csv")
df.to_csv(categorized_csv, index=False)
print(f"Archivo categorizado guardado como: {categorized_csv}")

return df
```

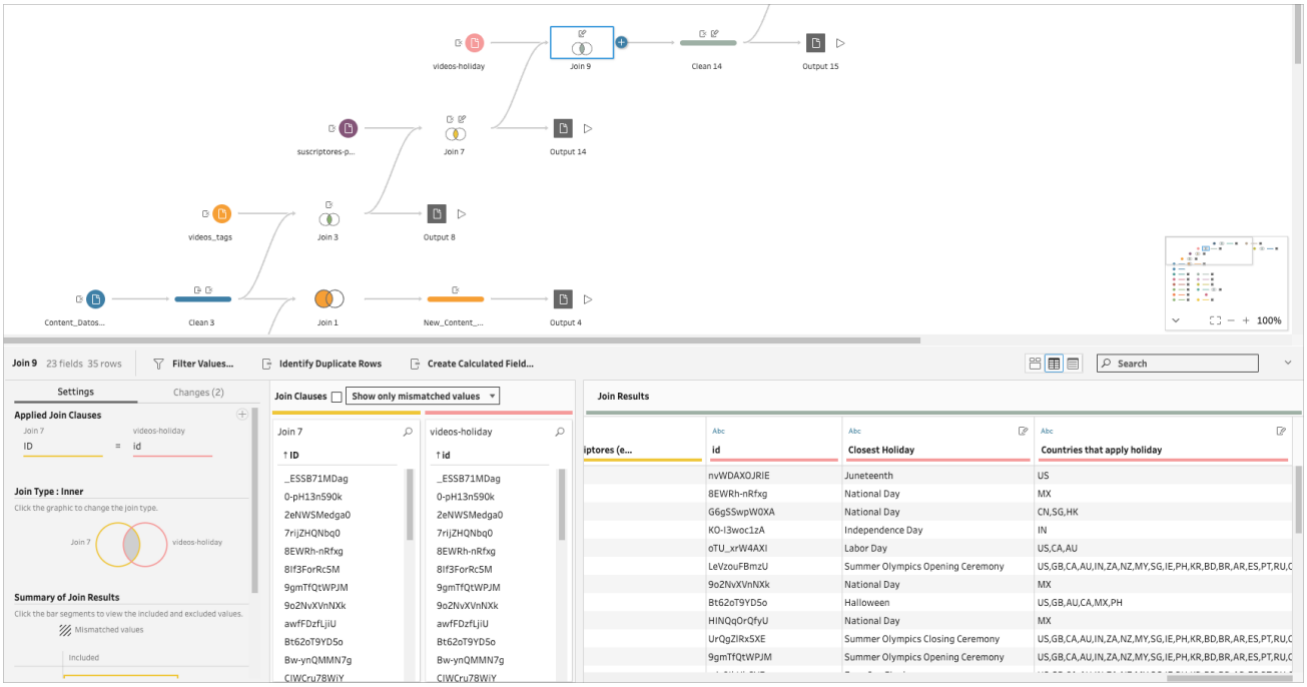
La segunda transformación, consiste en agregarle a los videos la fecha “limite” a la cual le atribuimos las estadísticas obtenidas desde que se publicó el video hasta que saliese un video nuevo, del mismo modo, agregar cuantos suscriptores se obtuvieron entre fechas y poder analizar un poco mejor el impacto de cada video dado un rango de fechas que vamos a considerar efectivas y que puede tener sentido asociar el video a dicho rango de fechas.



Del mismo modo, para obtener esa fecha, optamos por hacer un proceso en Python, para crear un dataset que contenga el ID del video y las fechas e informaciones adicionales que necesitábamos para poder comenzar a conectar la data de manera implícita asumiendo lo previamente mencionado.

```python
for i in range(len(videos_df)):
    # fecha de publicación del video actual
    current_date = videos_df.loc[i, 'Tiempo de publicación del video']
    video_id = videos_df.loc[i, 'ID']

    # fecha límite (día antes del próximo video o el máximo si es el último video)
    if i < len(videos_df) - 1:
        next_date = videos_df.loc[i + 1, 'Tiempo de publicación del video']
        limit_date = next_date - pd.Timedelta(days=1)
    else:
        limit_date = date_df['Fecha'].max()

    # filtrar suscriptores entre las fechas actuales
    mask = (date_df['Fecha'] >= current_date) & (date_df['Fecha'] <= limit_date)
    subscribers_count = date_df.loc[mask, 'Suscriptores'].sum()

    results.append({
        'id': video_id,
        'fecha_publicacion': current_date.date(),    "publicacion": Unknown word.
        'fecha_limite': limit_date.date(),
        'suscriptores_entre_fechas': subscribers_count
    })

results_df = pd.DataFrame(results)
results_df.to_csv('suscriptores_por_video.csv', index=False)
```

La tercera transformación consistió en relacionar los videos con algunas festividades que se celebran por lo menos en Estados Unidos y otras como las Olimpiadas o la Euro Copa o Halloween que son eventos o celebraciones conocidos a una escala mayor, del mismo modo colocamos en que países se celebra dicho evento o celebración.



Para obtener estas festividades, consultamos en internet algún listado sencillo con fechas y lo integramos en un CSV usando Python

```python
def find_closest_holiday(video_date, holidays_df):
    holidays_df['difference'] = abs(holidays_df['date'] - video_date)
    closest_holiday = holidays_df.loc[holidays_df['difference'].idxmin()]    "idxmin": Unknown word.
    return closest_holiday['holiday'], closest_holiday['countries']
```

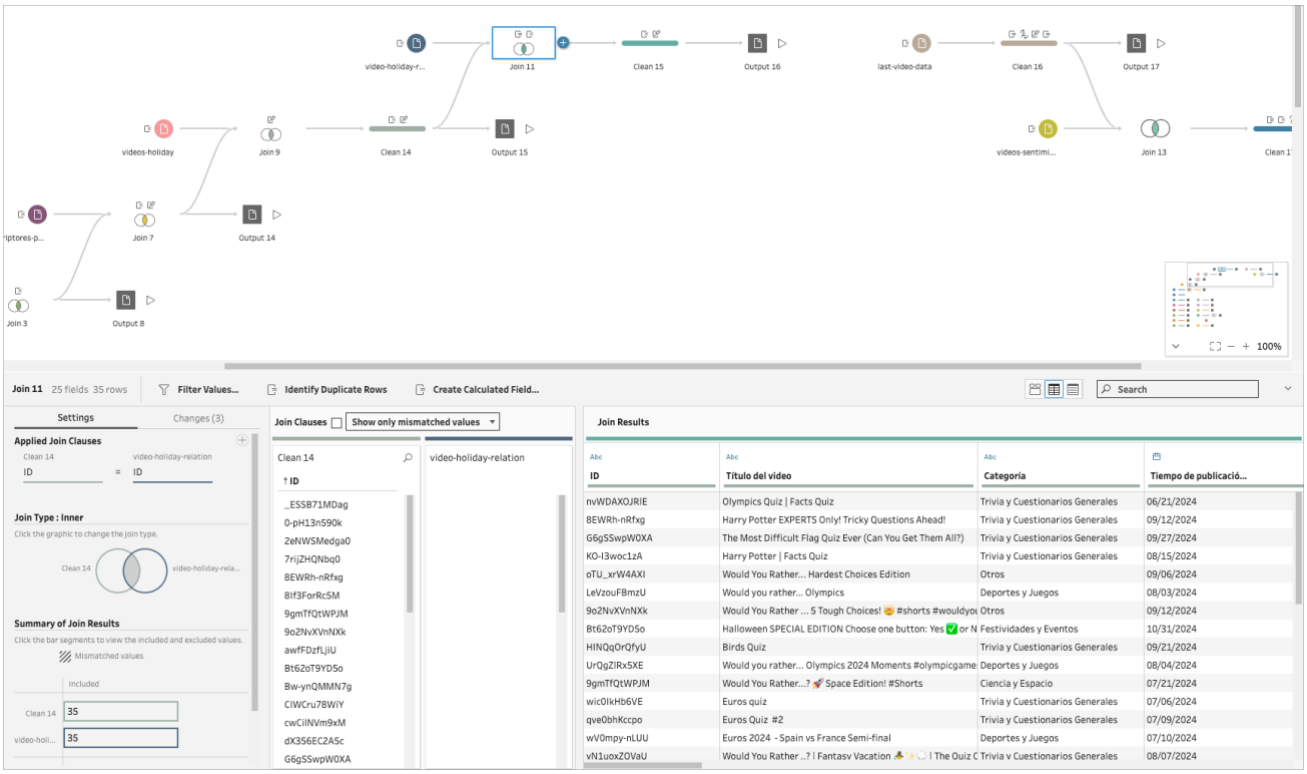
```python
video_holidays = []
for _, row in videos_df.iterrows():    "iterrows": Unknown word.
    holiday, countries = find_closest_holiday(row['Tiempo de publicación del video'], holidays_df)
    video_holidays.append({
        "id": row["ID"],
        "holiday": holiday,
        "countries": countries
    })
```

```python
video_holidays_df = pd.DataFrame(video_holidays)
```

```python
video_holidays_df.to_csv('videos-holiday.csv', index=False)
```

```python
videos_new_df = pd.read_csv('videos.csv')
```

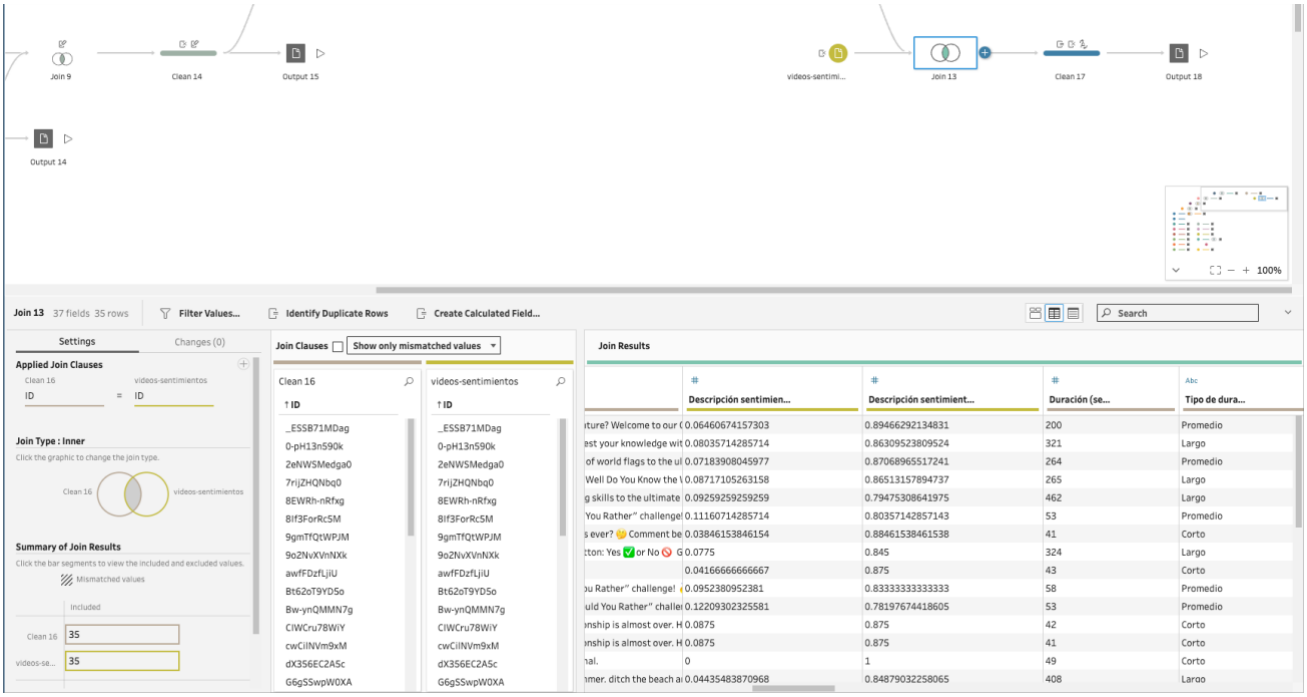
La cuarta transformación, corresponde a vincular si la información pasada, tiene algo que ver realmente con el video, entiéndase, si el contenido del video se relaciona con la festividad.



Para lograr esto, volvemos a Python y utilizamos técnicas de NLP para tratar de entender el contenido del video e integrar de manera automática si la festividad está relacionada.

```
def check_holiday_in_text(row):
    holiday = str(row['Celebración cercana']).lower()
    title = str(row['Título del video']).lower()
    description = str(row['Descripción']).lower()

    stemmer_en = SnowballStemmer('english')
    stemmer_es = SnowballStemmer('spanish')    "spanish": Unknown word.

    def stem_word(word):
        stems = set()
        stems.add(stemmer_en.stem(word))
        stems.add(stemmer_es.stem(word))
        return stems

    holiday_tokens = word_tokenize(holiday)
    stop_words = set(stopwords.words('english')) | set(stopwords.words('spanish'))    "stopwords": Unknown word.
    holiday_keywords = set()
    for word in holiday_tokens:
        if word.isalpha() and word not in stop_words:
            holiday_keywords.update(stem_word(word))

    text_tokens = word_tokenize(title) + word_tokenize(description)
    text_stems = set()
    for word in text_tokens:
        if word.isalpha():
            text_stems.update(stem_word(word))

    if holiday_keywords & text_stems:
        return True
    return False
```

La quinta transformación consistió en integrar el valor sentimental que aportaban las descripciones de los videos. Para esto volvimos a utilizar técnicas de NLP en Python.



En Python utilizamos Sentiwordnet para poder hacer los calculos correspondientes:

```python
df['Descripción Limpia'] = df['Descripción'].apply(limpiar_texto)
```

Python

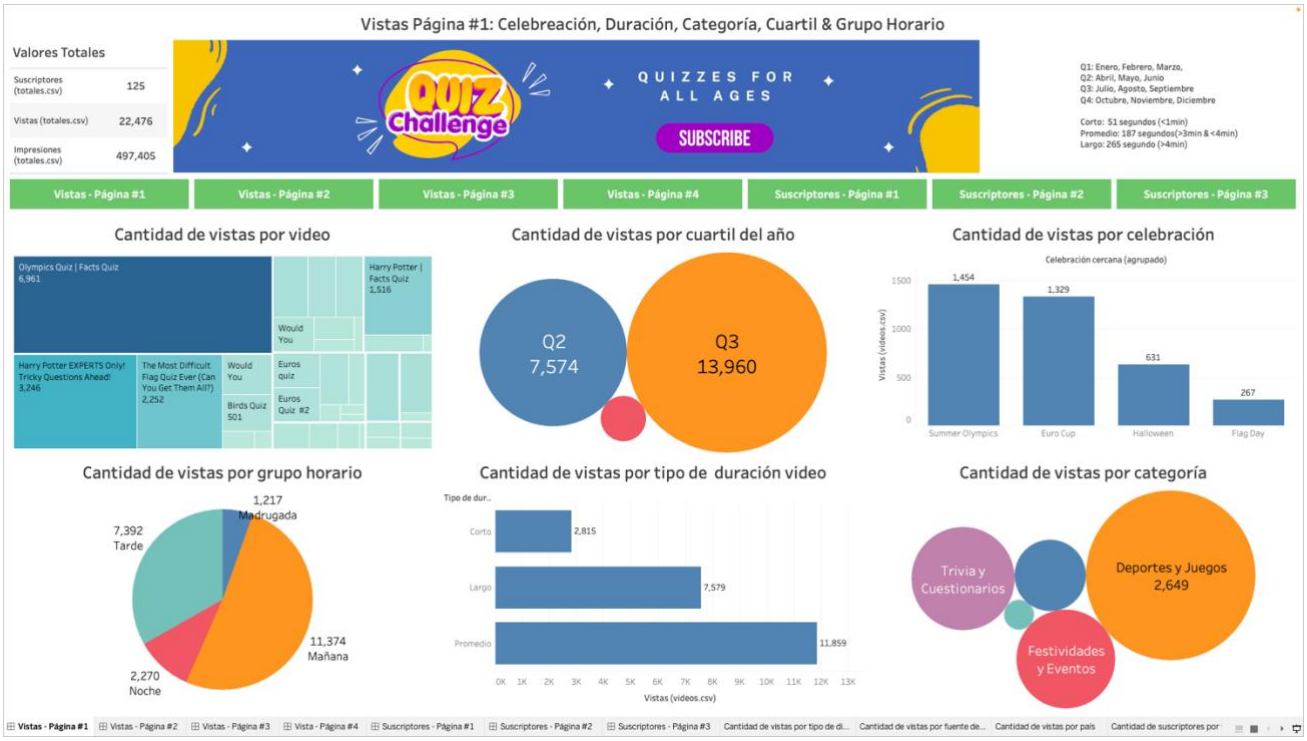
```python
def calcular_sentimiento(tokens):
    suma_p = 0.0
    suma_n = 0.0
    suma_s = 0.0
    contador = 0

    for palabra in tokens:
        synsets = list(swn.senti_synsets(palabra))    "synsets": Unknown word.
        if synsets:    "synsets": Unknown word.
            # tomamos el primer synset como representación principal     "synset": Unknown word.
            synset = synsets[0]    "synset": Unknown word.
            suma_p += synset.pos_score()    "synset": Unknown word.
            suma_n += synset.neg_score()    "synset": Unknown word.
            suma_s += synset.obj_score()    "synset": Unknown word.
            contador += 1

    if contador > 0:
        promedio_p = suma_p / contador
        promedio_n = suma_n / contador
        promedio_s = suma_s / contador
    else:
        promedio_p = 0.0
        promedio_n = 0.0
        promedio_s = 0.0

    return pd.Series([promedio_p, promedio_n, promedio_s])
```

Python

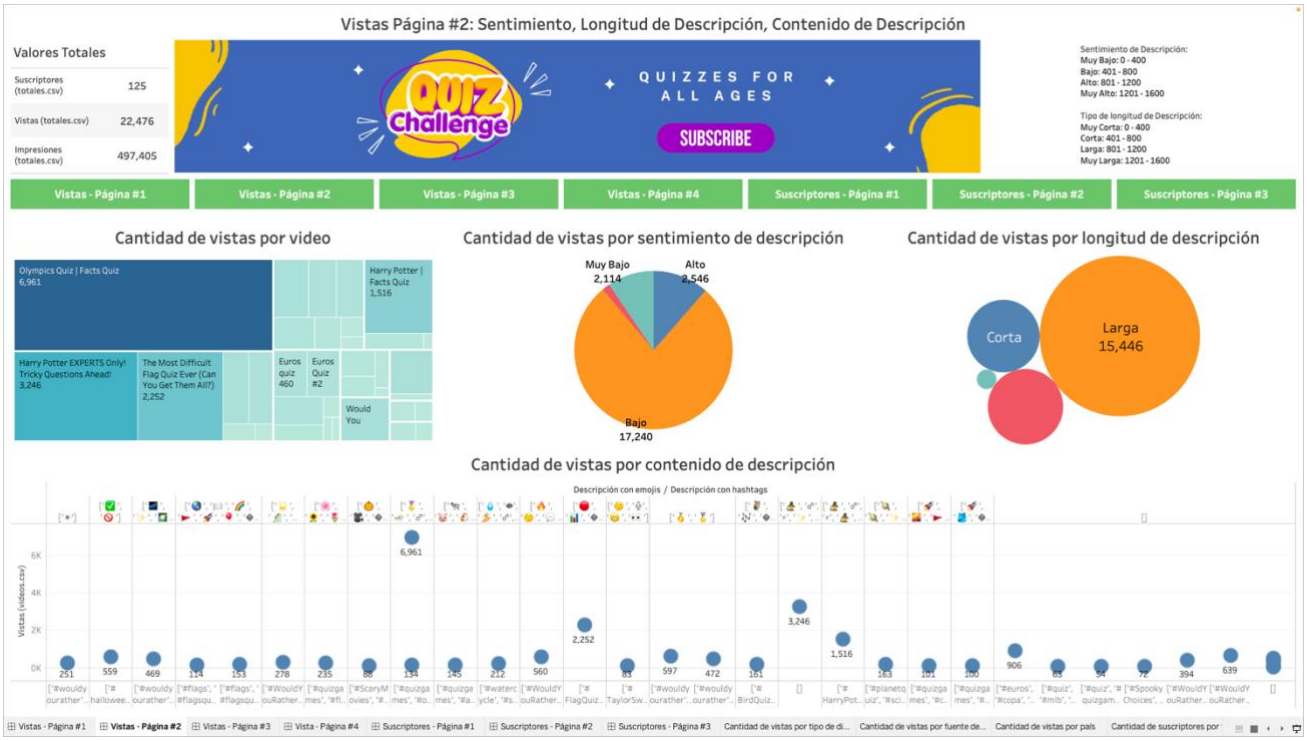
Con esto completamos la etapa de transformación de la información.

Dashboard

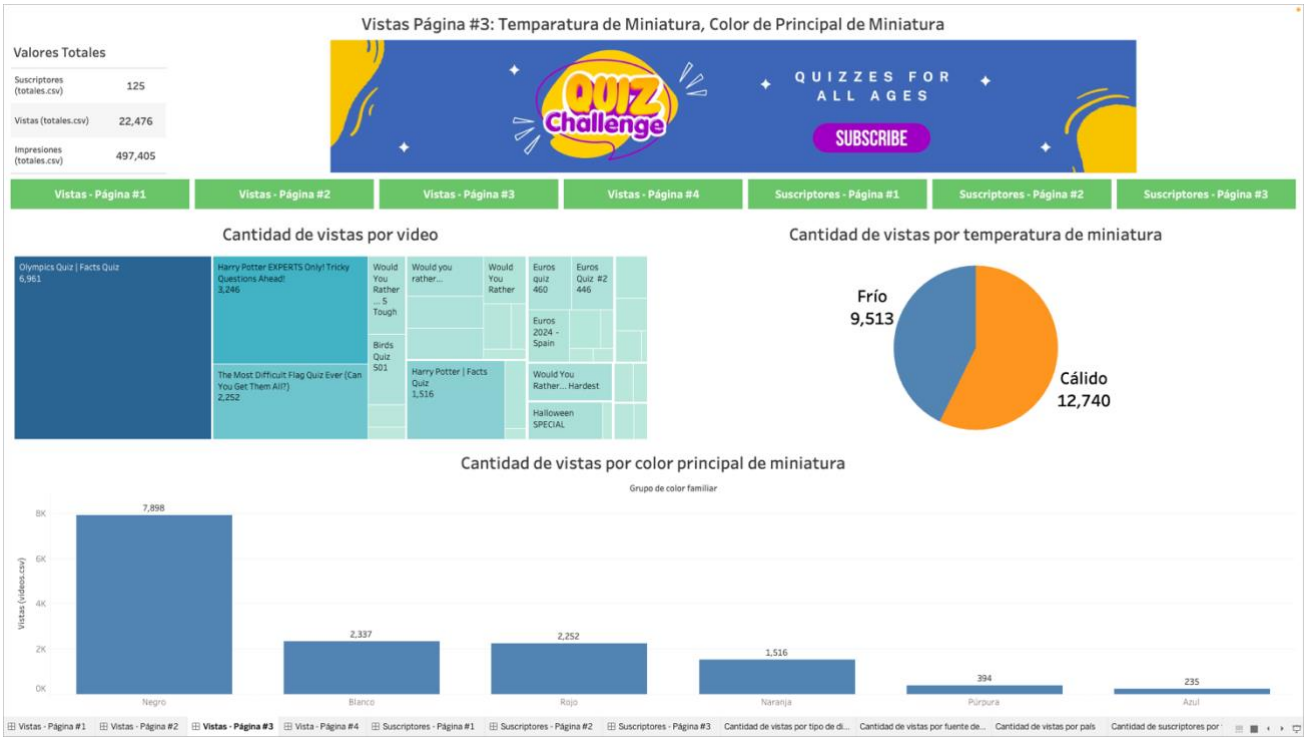
Una vez con toda la data lista completamente, realizar el dashboard como tal, no tuvo muchas complicaciones, fue más que nada la toma de decisión sobre qué información mostrar, como mostrarla y orquestar todos los filtros que logramos colocar. Cabe destacar que, de manera general, los datos están interconectados entre ellos y posible realizar cadenas de filtros para llegar a un resultado bien especifico.



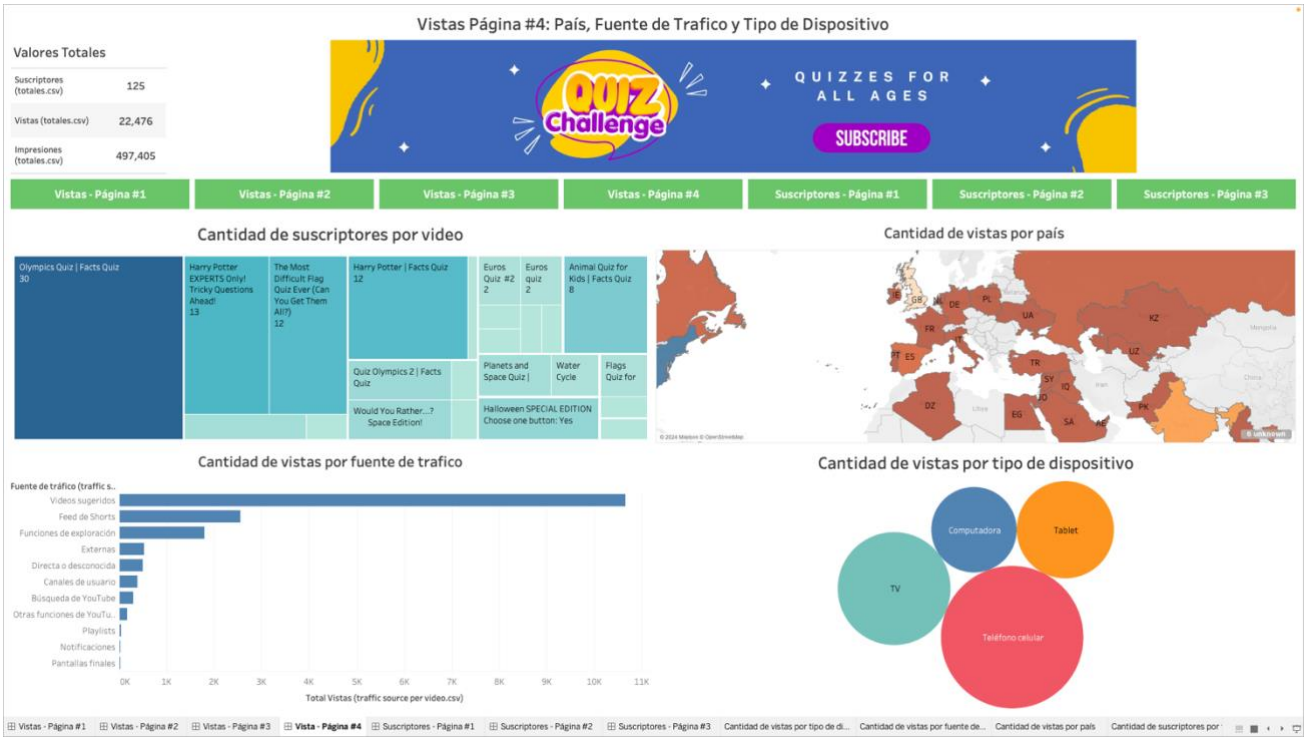
De esta primera página, podemos obtener datos enfocados en la fuente de vistas que el canal tiene, entiéndase cantidad de vistas por: video, cuartil del año, celebración relacionada (esto asegurando que el video sí está relacionado con dicho evento o celebración), grupo horario, duración de videos y la categoría de los mismo.



De esta segunda página, podemos obtener datos enfocados en la fuente de vistas que el canal tiene, entiéndase cantidad de vistas por: video, sentimiento de la descripción, longitud de la descripción y el contenido de la descripción

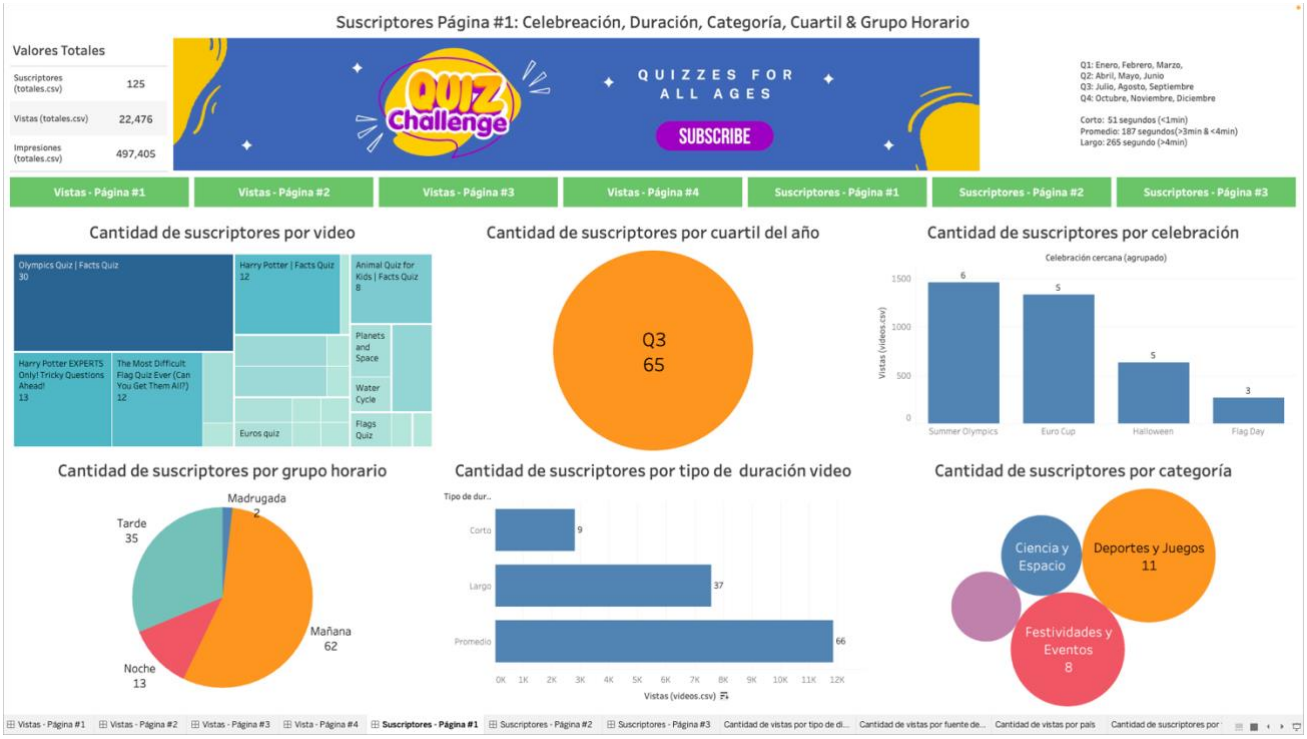


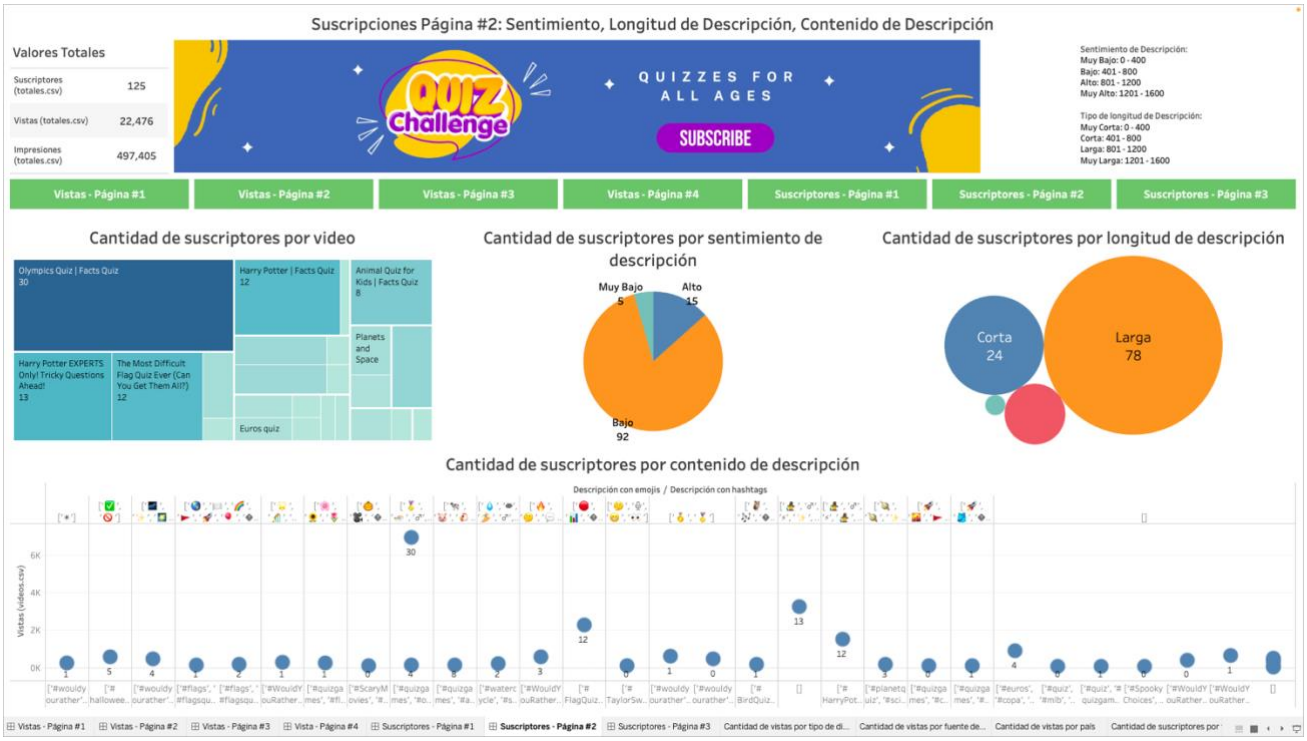
De esta tercera página, podemos obtener datos enfocados en la fuente de vistas que el canal tiene, entiéndase cantidad de vistas por: color principal de la miniatura y la temperatura de color que esta representa.

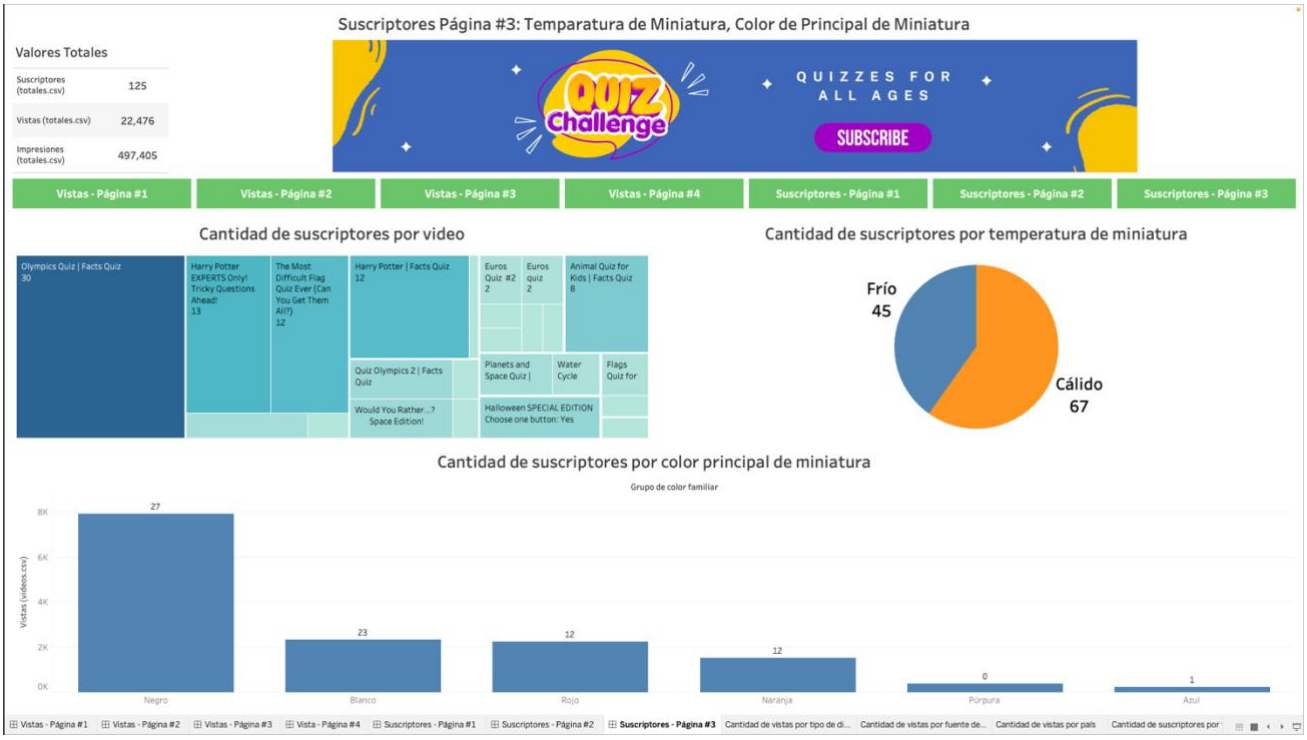


De esta cuarta y última página de vistas, podemos obtener datos enfocados en la fuente de vistas que el canal tiene, entiéndase cantidad de vistas por: País, Tipo de Dispositivo y Fuente de Trafico.

Para los suscriptores, tenemos las primeras 3 páginas de las vistas, pero adaptadas a las métricas de suscriptores:






Suscriptores Página #3: Temparatura de Miniatura, Color de Principal de Miniatura
Valores Totales
Suscriptores (totales.csv) 125
Vistas (totales.csv) 22,476
Impresiones (totales.csv) 497,405
QUIZZES FOR ALL AGES
SUBSCRIBE
Vistas - Página #1
Vistas - Página #2
Vistas - Página #3
Vistas - Página #4
Suscriptores - Página #1
Suscriptores - Página #2
Suscriptores - Página #3
Cantidad de suscriptores por video
Olympics Quiz | Facts Quiz 30
Harry Potter EXPERTS Only! Tricky Questions Ahead! 13
The Most Difficult Flag Quiz Ever (Can You Get Them All?) 12
Harry Potter | Facts Quiz 12
Euros Quiz #2 2
Euros quiz 2
Animal Quiz for Kids | Facts Quiz 8
Quiz Olympics 2 | Facts Quiz
Planets and Space Quiz |
Water Cycle
Flags Quiz for
Would You Rather...? Space Edition!
Halloween SPECIAL EDITION Choose one button: Yes
Cantidad de suscriptores por temperatura de miniatura
Frío 45
Cálido 67
Cantidad de suscriptores por color principal de miniatura
Grupo de color familiar
Vistas (videos.csv)
8K
6K
4K
2K
0K
27
23
12
12
0
1
Negro
Blanco
Rojo
Naranja
Púrpura
Azul
Vistas - Página #1
Vistas - Página #2
Vistas - Página #3
Vista - Página #4
Suscriptores - Página #1
Suscriptores - Página #2
Suscriptores - Página #3
Cantidad de vistas por tipo de di...
Cantidad de vistas por fuente de...
Cantidad de vistas por país
Cantidad de suscriptores por

Conclusión

El desarrollo de este dashboard interactivo permite a *The Quiz Challenge* acceder a una herramienta poderosa para comprender y analizar en profundidad las métricas clave de su canal, como vistas y suscriptores. Gracias al proceso de transformación y enriquecimiento de los datos, el dashboard ofrece una visión detallada y personalizada que abarca múltiples perspectivas del contenido y su impacto en la audiencia.

Algunas de las principales ventajas que brinda el dashboard incluyen:

1. **Identificación de patrones y tendencias**: Se pueden analizar factores como la duración de los videos, la relación con eventos y celebraciones, el contenido de las descripciones y títulos, el uso de emojis y hashtags, y cómo estos elementos influyen en el desempeño del canal.

2. **Exploración de aspectos visuales**: El análisis del color predominante en las miniaturas y su relación con el engagement es un aporte único que ayuda a entender cómo los aspectos visuales pueden influir en las vistas.

3. **Impacto de los eventos externos**: Al vincular videos con festividades y eventos globales, se puede determinar qué tan relevante es el contenido para aprovechar tendencias específicas y aumentar el interés de la audiencia.

4. **Segmentación por países y dispositivos**: Los datos geográficos y de tipo de dispositivo ofrecen una mejor comprensión del público objetivo, ayudando a personalizar estrategias para diferentes segmentos.

5. **Relación entre sentimientos y resultados**: La incorporación de análisis de sentimiento en las descripciones permite medir el impacto emocional del contenido en la audiencia, ayudando a crear contenido más atractivo.

En conjunto, el dashboard no solo ofrece respuestas inmediatas a las preguntas del canal, sino que también abre puertas a estrategias basadas en datos que facilitan la toma de decisiones informadas. Esto constituye un paso significativo hacia la mejora del engagement con la audiencia y el logro de objetivos como la monetización del canal y su crecimiento sostenible.